Person Foreground Segmentation by Learning Multi-domain Networks

Zhiyuan Liang, Kan Guo, Xiaobo Li, Xiaogang Jin, *Member, IEEE*, and Jianbing Shen, *Senior Member, IEEE*

***Abstract*—Separating the dominant person from the complex background is significant to the human-related research and photo-editing based applications. Existing segmentation algorithms are either too general to separate the person region accurately, or not capable of achieving real-time speed. In this paper, we introduce the multi-domain learning framework into a novel baseline model to construct the Multi-domain TriSeNet Networks for the real-time single person image segmentation. We first divide training data into different subdomains based on the characteristics of single person images, then apply a multi-branch Feature Fusion Module (FFM) to decouple the networks into the domain-independent and the domain-specific layers. To further enhance the accuracy, a self-supervised learning strategy is proposed to dig out domain relations during training. It helps transfer domain-specific knowledge by improving predictive consistency among different FFM branches. Moreover, we create a large-scale single person image segmentation dataset named MSSP20k, which consists of 22,100 pixel-level annotated images in the real world. The MSSP20k dataset is more complex and challenging than existing public ones in terms of scalability and variety, which benefits much to human segmentation research. Experiments show that our Multi-domain TriSeNet outperforms state-of-the-art approaches on both public and the newly built datasets with real-time speed.**

***Index Terms*—single person segmentation, light-weight networks, multi-domain learning.**

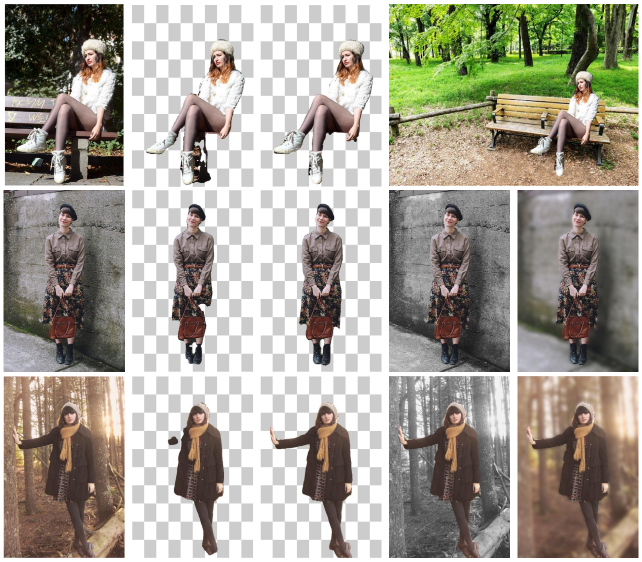


Fig. 1. Examples of the photo-editing applications benefiting from the single person segmentation. From left to right, column (a) shows input images, (b) shows segmentation results of BiSeNet [36], (c) shows the results of our approach, and (d) shows the edited images based on our segmentation results.

I. Introduction

Real-time single person segmentation aims to understand the human content of the images and separate the dominant person under resource constraints. It is the foundation of many human-based photo-editing applications as shown in Fig. 1, and can stimulate the human-related computer vision research such as person re-identification [39], [27], human behavior analysis [10], and multiple object tracking [14], [32].

Recently, Convolutional Neural Networks (CNNs) have witnessed the success in human-related image segmentation for its power of feature representations. Wu *et al.* [33] first present a CNNs-based single person segmentation method to extract hierarchical context information from multi-scale inputs. Song *et al.* [31] discard the time-consuming multiple input process and propose a faster method by adopting the light-weight backbone and reducing the resolution of the input image. The outputs of these two methods are generated through several fully-connected layers, resulting in the loss of spatial information. To generate more accurate segmentation predictions, some specialized algorithms are proposed to handle the single person segmentation problem under specific image domains. Shen *et al.* [30] propose an automatic portrait segmentation algorithm. They leverage the domain-specific knowledge and extend the FCN framework [22] by introducing the portrait position and shape channels. Chen *et al.* [6] design a new loss function based on the image-level gradient information to generate sharper boundary of the portraits. Zhu *et al.* [43] propose a light-weight deep matting algorithm for portrait animation by adding the dilated convolution into the dense block structure. However, these portrait segmentation and matting algorithms only carry out the half-body segmentation task and do not consider the full-body or far-shot scenarios, which restricts their applications. In this paper, we aim to design a real-time single person segmentation method that can separate the dominant person under different image scenarios.

We present a novel baseline model denoted as TriSeNet for real-time single person segmentation. The TriSeNet takes full advantage of the feature representation of CNNs. It

Z. Liang is with Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, P. R. China, and also with Alibaba Group.

K. Guo and X. Li are with Alibaba Group, Hangzhou 311121, P. R. China. (email: guokan.gk and xiaobo.lixb@alibaba-inc.com)

X. Jin is with State Key Lab of CAD&CG, Zhejiang University, Hangzhou 310058, P. R. China. (email: jin@cad.zju.edu.cn)

J. Shen is with Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, Beijing 100081, P. R. China. (email: shenjianbingcg@gmail.com) (Corresponding authors: Xiaogang Jin and Jianbing Shen.)

first extracts the high-resolution spatial features, high-level semantic features, and detailed body boundary features, then fuses them jointly with a Feature Fusion Module (FFM). However, it is not easy to learn a unified representation since the segmentation algorithms suffer from different challenges when the distance from the dominant person to the camera lens varies. For example, headshots often have a simpler background but with differences in the foreground such as the hairline. Far-shot images contain fewer human details while the various clothing and the cluttering background should be carefully considered. The intra-domain gap still exists within a single segmentation dataset [42]. Due to the variation and inconsistency across single person images, we divide them into different subdomains and propose a multi-domain learning framework to tackle this problem. In detail, we detect the face and compute the ratio of face area in each image. The images are classified into six subcategories according to the ratio and each subcategory is regarded as a separate subdomain. A multi-branch FFM is introduced to decouple the networks into the domain-independent and the domain-specific layers. Besides, a self-supervised learning strategy is employed to improve knowledge sharing across different subdomains. By considering the prediction of other branches as the soft label, each FFM branch is not only trained with subdomain data directly but also supervised by the predictions of other subdomains indirectly. As a result, our method can improve the predictive consistency among different FFM branches and can obtain higher robustness under challenging conditions.

The dataset is another significant factor to hamper the development of the single person segmentation research. Existing datasets suffer from the monotonous background or are not available on the Internet. The public datasets like the Baidu people segmentation database [33] are composed of limited images, which constrains the potential to train deeper networks. In this paper, we build a large-scale single person segmentation dataset consisting of more than 20k **M**ulti-**S**cale **S**ingle **P**erson images from the real world. The newly built dataset is named as MSSP20k. It contains representative instances with various clothing, complex postures, cluttering backgrounds, partial occlusions, and a wide range of viewpoints. All the images are fine-labeled for the single person segmentation task. Compared to the dataset in [33], the MSSP20k dataset benefits the single person segmentation research better in terms of both data volume and appearance variety. The new dataset will be made publicly available. More details about this dataset are discussed in Section III.

The main contributions of our paper are summarized as follows:

- A novel TriSeNet architecture is proposed for the real-time single person segmentation task. It first extracts high-resolution spatial features, high-level semantic features, and detailed boundary features, then fuses them jointly with the Feature Fusion Module (FFM). The proposed networks can separate the foreground person with accurate body boundaries.
- A multi-domain learning framework is proposed based on our TriSeNet. The training data are divided into different subdomains and a multi-branch FFM is proposed to learn both domain-independent and domain-specific representations while maintaining a high speed. Moreover, a self-supervised learning strategy is proposed to further enhance the performance by improving the predictive consistency across different FFM branches.
- We create a large-scale dataset, named as MSSP20k for real-world single person image segmentation. The dataset, which will be made publicly available, contains 22,100 high-quality images with pixel-level annotations. The MSSP20k dataset not only makes it possible to train deep networks directly for the single person segmentation task but also contributes to the human-related research for its volume and diversity.
- Our method achieves state-of-the-art performance on the public and the newly built datasets. Specifically, it achieves 93.49% IoU on the Baidu people segmentation database and 91.52% IoU on the newly built MSSP20k dataset. Our method can be performed at 120 FPS on a single TESLA P100 GPU.

II. Related Works

A. Single Person Segmentation

Wu *et al.* [33] propose the first deep convolutional networks for single person segmentation. They utilize three feature extractors with different scales to capture hierarchical feature presentations. However, their method is time-consuming for multiple inputs. Song *et al.* [31] take into consideration both efficiency and effectiveness of deep models via reducing the resolution of input images and adopting deeper networks. Although their method can be performed at a high speed, the decreased accuracy prevents it from being widely used. Recently, some specialized algorithms have been studied to generate more accurate predictions for portrait images. Shen *et al.* [30] first address the problem of the automatic portrait segmentation by introducing the portrait position and shape input channels into the FCN framework. BANet [6] proposes to selectively extract the detailed information in boundary areas and designed a new loss based on the image-level gradient information. PortraitNet [37] constructs a light-weight U-shape framework with two auxiliary losses, which preserve boundary details and improve the consistency between different illumination images. Since portrait images only depict face, hair, and shoulders, the portrait shape and the background are relatively simple. These algorithms can achieve good performance even using a straightforward end-to-end architecture. However, the single person segmentation task is not only limited to portraits but also a variety of images in other domains (e.g. full-body and far-shot images). In this paper, we present a real-time single person segmentation method, which can carry out the segmentation task under different conditions and achieve better performance at a high speed.

B. Real-time Image Segmentation

Some image segmentation algorithms pay more attention to computational efficiency by introducing light-weight convolutional networks. ENet [24] utilizes the bottleneck structure

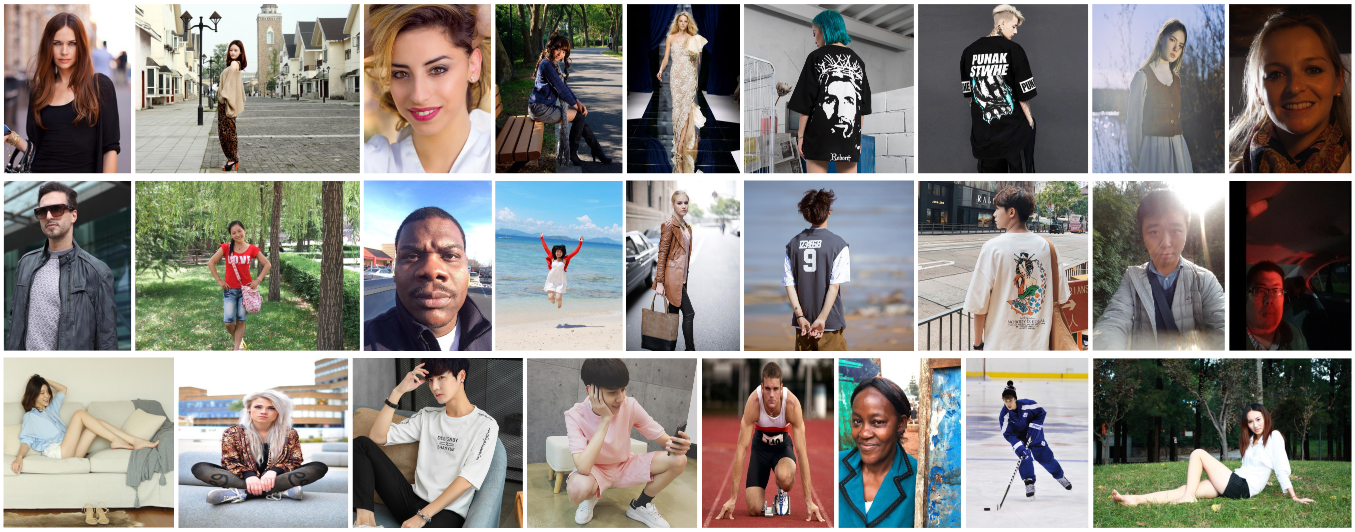


Fig. 2. Some representative samples in the MSSP20k dataset. The images have a variety of resolutions and contain only one dominant person. Our dataset has a large variety of pose, clothing, hairstyle, skin colors, focal length, illumination conditions, and background.

of ResNet and reduces the channel dimension of features. It achieves much faster speed by sacrificing accuracy. SegNet [1] adopts the light-weight network architecture with the skip connection. The resolution of feature representations is decreased with an encoder and restored with a decoder, and thus both the low-level and high-level information of the image are aggregated. DLC [18] presents a real-time semantic segmentation method. The unconfident regions are forward propagated in the cascade networks and the computational complexity of easy regions is reduced. ICNet [38] first uses three streams to extract features from images of different resolutions, then aggregates them with the cascaded feature fusion unit. ESPNetv2 [49] designs an efficient network with group point-wise and depth-wise dilated separable convolutions. DFANet [50] stacks the light-weight backbones multiple times to aggregate discriminative features through subnetwork and sub-stage cascade, respectively. BiSeNet [36] is the most related semantic segmentation algorithm to our baseline model. It utilizes a two-stream framework to extract spatial and context information respectively and then fuses them with the Feature Fusion Module (FFM). In this paper, we introduce another edge path into BiSeNet and propose the TriSeNet architecture for the real-time single person segmentation. The proposed method can separate the dominant person with accurate human boundaries with real-time speed. To further improve the performance of TriSeNet, we propose a multi-domain learning framework to construct the Multi-domain TriSeNet. It first classifies single person images into six subcategories according to the scale of the face area, then selects a proper FFM branch to carry out segmentation. A self-supervised learning strategy is further applied to the Multi-domain TriSeNet during the training phase, which contributes to knowledge sharing across different subdomains without requiring additional computational cost during the testing phase. The proposed method can achieve state-of-the-art performance at 120 FPS. Due to the computational efficiency of our method, it can be easily extended to the video segmentation methods by leveraging the memory modules [51] or multi-scale spatial-temporal contexts [52] in future work.

C. Multi-domain Learning

Multi-domain Learning (MDL) aims to learn a model with minimal average risk across multi-domain data [46]. MDL has been applied in many computer vision research including object classification, person re-identification, and visual object tracking. Hoffman *et al.* [45] designed a hierarchical clustering method to discover latent domains and proposed a domain transform mixture model for object classification. Xiao *et al.* [48] propose the domain-guided dropout to learn robust feature representations from multiple domains. Yang *et al.* [47] unify multi-domain learning and multi-task learning into one framework via the prior knowledge of domain semantic relationships. MDNet [44] is the most related algorithm to our MDL framework. MDNet regards each video sequence as a separate domain. The parameters of shared layers are learned during training and the parameters of classification layers are fine-tuned by online updating. The main difference between MDNet and ours is twofold. First, there is no explicit dividing standard for multiple domains in MDNet, so it discarded domain-specific layers at the inference time. Our method quantifies the standard which can be reused during testing, thus both the domain-independent and the domain-specific knowledge are preserved. Second, MDNet requires online updating to adjust new inputs while our method replaces the time-consuming online updating process with high-speed face detection, which costs negligible computational overhead. Compare to MDNet, our method can take full advantage of MDL while maintaining a high speed.

III. MSSP20k Dataset

Although there have been many human-related segmentation datasets now, they can hardly be applied to single person seg-

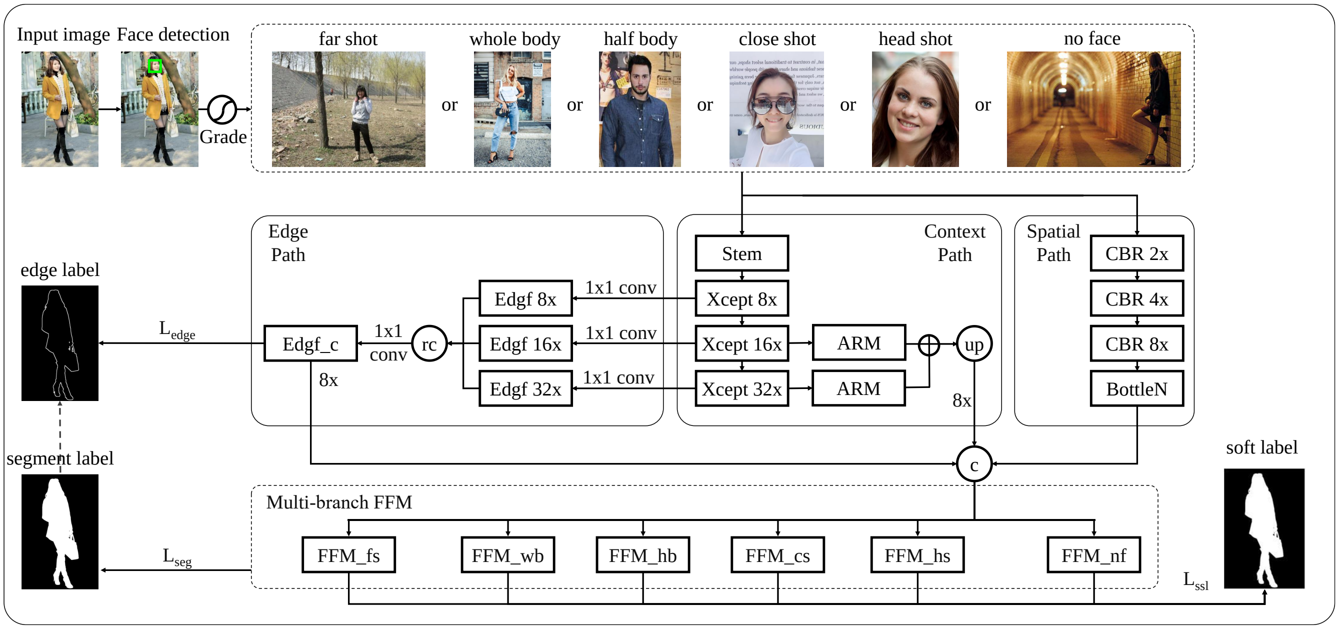


Fig. 3. The pipeline of our approach. Given a single person image, we detect the face of the dominant person and classify the image into the proper subdomain. Then the image is fed into the network to extract spatial features, context features, and edge features, respectively. ARM is the Attention Refinement Module proposed in BiSeNet. FFM is the Feature Fusion Module. During the training step, these three types of features are fused into all the FFM branches to learn both the domain-independent and the domain-specific representations. During the test step, a proper FFM branch is selected based on the face-assisted classification and more accurate predictions can be generated. In the pipeline, "c" means concatenating operation. "rc" means rescaling and concatenating operations, and "up" means upsampling operation with bilinear interpolation. L_{seg}, L_{edge}, and L_{ssl} are segmentation loss, edge loss, and self-supervised learning loss, respectively.

mentation directly. For example, human parsing datasets [11], [19] aim to separate human bodies into different semantic parts, so they focus more on foreground objects and ignore the background. The instances in the datasets tend to be close to the image boundary but the far-shot images with the various background are insufficient. Portrait segmentation datasets [30] are built to separate the upper bodies from the background, but the datasets are too specialized to cover all situations in single person segmentation problem. SHM [41] proposed a human matting dataset. However, the images in the dataset are synthetic so they are not suitable for single person segmentation in the real world. The single person segmentation is fundamental research for many tasks. The only publicly available dataset for this research is Baidu people segmentation database [33], which contains about 6,700 labeled single person images from the Internet. As a result, existing public datasets are far from the requirement for the single person segmentation research.

In this paper, we collect and build a high-quality single person segmentation dataset. All images come from real-world scenes. Part of the collected images are from the Internet, and the rest are from daily photos. We collect more than 100k images and have a fine screening to eliminate easy ones such as a background with pure color. We ensure that each image only contains one dominant instance. Instances with a small area are considered as background. The foreground area includes the body (e.g. arms and legs), clothing (e.g. dresses and hats), and any objects in the hands (e.g. handbags and telephones) of the dominant person. Some representative examples in the MSSP20k dataset are shown in Fig. 2. The images in our new dataset have larger variations in pose, clothing, hairstyle, focal length, background, and so on. Table I shows the statistics in comparison to the Baidu people segmentation database [33]. We use RH to represent the Ratio of Half-body images in the dataset. A half-body image only contains the upper body of the dominant person, and the legs or feet will not appear. Compared to the Baidu database, our MSSP20k has a more balanced data distribution between the half-body and the full-body images, which demonstrates the variety of foreground instances. We use RI to represent the Ratio of Indoor images in the dataset, which measures the variety of background scenes. As shown in Table I, the Baidu database focuses more on outdoor scenes while our MSSP20k dataset focuses more on indoor scenes. We use MR to represent the Mean Resolution of images. The Baidu database is larger than ours while our MSSP20k has a larger total number of images. The Baidu database consists of 5,387 training images and 1,316 testing images whereas our newly built MSSP20k has a larger volume in terms of both training and testing sets. The MSSP20k dataset covers a large number of single person images in real-world scenes with different resolutions range from 140×140 to 3000×3000. We collect $22,100$ images in total, which are four times the volume of the Baidu database. All of the images are fine-labeled by professional labeling workers. The dominant person in the image is labeled as foreground and other pixels are labeled as background. We split the labeled images randomly into a training set with $18,000$ images and a testing set with $4,100$ images.

TABLE I
OVERVIEW OF THE PUBLIC DATASETS FOR SINGLE PERSON SEGMENTATION. WE USE RH, RI AND MR TO REPRESENT THE RATIO OF HALF-BODY IMAGES, THE RATIO OF INDOOR IMAGES, AND THE MEAN RESOLUTION OF IMAGES IN THE DATASET, RESPECTIVELY. WE ALSO REPORT THE NUMBER OF ANNOTATED IMAGES FOR TRAINING AND TESTING SETS.

Dataset	RH	RI	MR	Training	Testing	Total
Baidu database [33]	0.01	0.34	1,053×803	5,387	1,316	6,703
MSSP20k	0.54	0.62	913×765	18,000	4,100	22,100

IV. OUR APPROACH

An overview of our method is shown in Fig. 3. We first elaborate on our baseline model TriSeNet (§ IV-A). Then we propose the multi-domain learning framework and extend the baseline model with a multi-branch FFM to construct our Multi-domain TriSeNet (§ IV-B). Moreover, we propose a self-supervised learning strategy (§ IV-C) to jointly training the multi-branch networks, which contributes to knowledge sharing by improving the predictive consistency among different FFM branches during training.

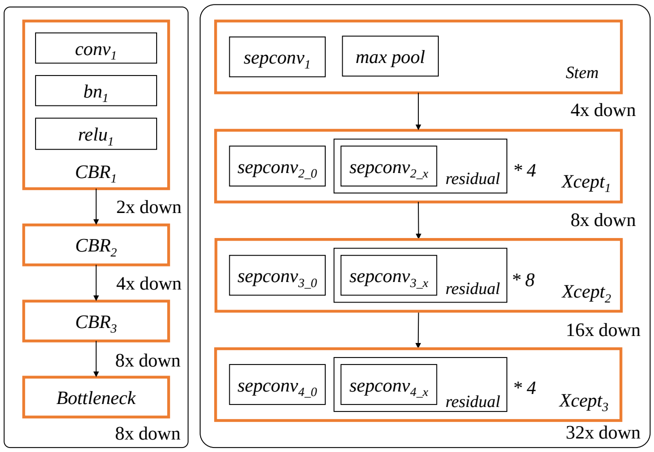
A. TriSeNet for Single Person Segmentation

Considering both accuracy and speed, BiSeNet [36] is employed as our basic model, which consists of two kinds of convolutional paths. The architecture of the two paths is shown in Fig. 4. The spatial path has three CBR blocks and one Bottleneck block to extract high-resolution spatial features. Each CBR block contains a 3×3 convolution layer with $stride = 2$, followed by a Batch Normalization layer and a ReLU activation layer. The Bottleneck block replaces the 3×3 convolution layer with a 1×1 convolution layer to increase the channel dimension of output features. Therefore, the resolution of output features of the spatial path is $1/8$ of the original image. The context path is much deeper than the spatial one. Taking Xception39 [7] as an example, the context path consists of one stem network and three Xception blocks. As shown in Fig. 3, to enhance the semantic features, the outputs of the last two blocks are first fed into the Attention Refinement Module (ARM) [36]. The refined features are gathered with a U-shape structure [28], which increases the resolution from $1/32$ to $1/8$ of the input image and gathers the semantic context with a wider range of receptive fields. Finally, the spatial features and the context features are fused by the Feature Fusion Module. The pixel-wise cross-entropy loss for the binary segmentation is employed:

$$\begin{aligned} L_{seg} &= -\sum_i y_i \log P(y_i=1|X) + (1-y_i)\log\left(1-P(y_i=1|X)\right) \\ &= -\sum_{i\in Y_+} \log P(y_i=1|X) - \sum_{i\in Y_-} \log P(y_i=0|X) \end{aligned} \tag{1}$$

where $y_i \in \{0,1\}, i=0,...,|X|$ is the ground-truth label of the input image. Y_+ and Y_- are the subsets of positive and negative pixels of the ground-truth label, respectively. Moreover, we find that the structure of the spatial path is relatively simple, and it has the potential to be further improved.

Inspired by xUnit activation [16], [15], we propose an efficient activation layer denoted as modified xUnit (mxUnit). Given the same number of convolutional layers, mxUnit can extract more discriminative feature representations in comparison to ReLU activation. The network structures of them are shown in Fig. 5. The ReLU activation layer utilizes a threshold function to set a binary weight to every pixel. The output is the multiplication of the input value and the weight value. The threshold function can be formulated as follows:



(a) spatial path (b) context path

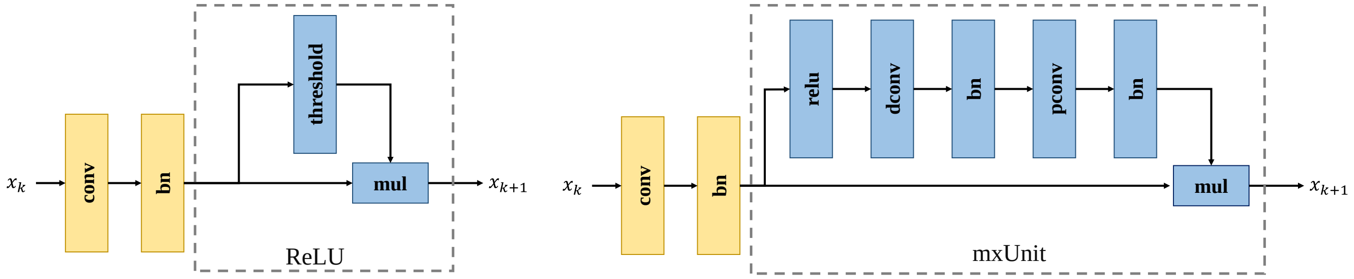
Fig. 4. The architecture of BiSeNet with the Xception39 backbone. (a) is the spatial path, which consists of three CBR blocks and one Bottleneck block. (b) is the context path, where sepconv means the depthwise separable convolution in [7], [13], [29]. It consists of one Stem block and three Xception blocks.

$$threshold([z_k]_i) = \begin{cases} 1 & [z_k]_i > 0 \\ 0 & [z_k]_i \leq 0 \end{cases} \tag{2}$$

where $[z_k]_i$ is the input of the k^{th} ReLU layer at location i. It is worth noting that the weight value of the pixel is only based on itself. Compared to non-parameterized ReLU activation, mxUnit is a parameterized activation which can enhance the feature representation using local features statics. The 5×5 depthwise convolution layer aggregates local activations in the spatial dimension and the pointwise convolution layer transfers the aggregations in the channel dimension. Followed by a Sigmoid layer, the weight value ranges from 0 to 1. In our implementations, we replace the ReLU activation layer with a mxUnit activation layer in the last two CBR blocks of the spatial path.

To further improve the performance, we add an edge path into BiSeNet and construct the TriSeNet architecture. As shown in Fig. 3, multi-scale semantic features are utilized. The edge loss is adopted during training:

$$L_{edge} = -\left[\beta \sum_{j\in Y_+} \log P(y_j=1|X) \; + (1-\beta) \sum_{j\in Y_-} \log P(y_j=0|X)\right] \tag{3}$$



(a) representation of ReLU activation

(b) our proposed mxUnit activation

Fig. 5. The architecture of (a) the ReLU activation layer and (b) the proposed modified xUnit (mxUnit) activation layer. The threshold function in ReLU generates a weight map with a binary value of {0,1}. The proposed mxUnit aggregates information from spatial and channel dimensions and generates a weight map in the range of [0,1] through a sigmoid operation.

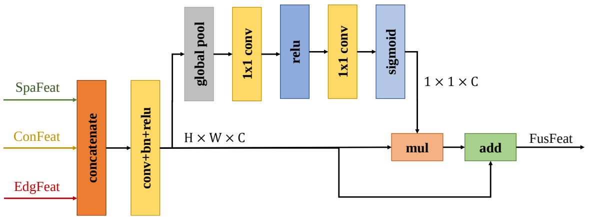


Fig. 6. The architecture of Feature Fusion Module (FFM) used in our method. SpaFeat, ConFeat, and EdgeFeat are features extracted by the spatial path, context path, and edge path, respectively. FusFeat is the fusion feature after the channel-wise multiplication and the residual connection.

where $\beta = |Y_-|/|Y|$ is a balance factor applied in imbalanced binary classification tasks. Different from BiSeNet, TriSeNet feeds high-resolution spatial features, high-level semantic features, and detailed body boundary features into the Feature Fusion Module, and generates predictions with a single person segmentation network head. The architecture of FFM used in our method is shown in Fig. 6. It can aggregate multiple semantic cues adaptively, which has been proved effective in many visual tasks, such as trajectory prediction [53] and saliency detection [54]. Inspired by the SE-Net [55], the FFM first combines three types of features with a convolutional layer, and then assigns channel weights for the combined features. A residual connection is adopted to preserve more details of input features. The total loss is a combination of the segmentation loss and edge loss:

$$L = L_{seg} + \lambda_1 L_{edge} \tag{4}$$

where λ_1 is the loss weight to balance the segmentation loss L_{seg} and the edge loss L_{edge}. The proposed TriSeNet can generate foreground masks with more accurate human boundaries in comparison to the basic model, which is important to photo-editing based applications.

B. Multi-domain TriSeNet with Face-assisted Classification

Multi-domain learning (MDL) aims at training a model across multiple domains, where each domain displays different experimental bias and the domain information is incorporated in the learning procedure. A challenge in MDL is how to best leverage information across multiple domains on the same subject, and how to dig out the knowledge that could not have been learned from any individual domain alone. The goal of our MDL framework is to train a multi-domain network separating the dominant person from the background under arbitrary image conditions. Our MDL framework consists of two steps. Firstly, the training data is classified into six subcategories based on the distance from the person to the lens. We regard each subcategory data as a subdomain. Secondly, a multi-branch FFM is developed in the baseline model to construct the Multi-domain TriSeNet, which facilitates learning both the domain-independent and the domain-specific feature representation.

The human photography has a certain regularity in the real world. We observe that the scale of the face is an important factor measuring the depth of the field. When the scale of the face is small, the depth of the field is large, and vice versa. As a result, we can use an automatic face detection method to grade the data. In our approach, the single person images are graded into six levels, they are named as $far\ shot$, $whole\ body$, $half\ body$, $close\ shot$, $head\ shot$, and $no\ face$. A fast face detection method [23] is employed. To achieve a higher speed, the inference framework is optimized so that it takes only a few milliseconds to detect a face. Then the ratio of face area in the image is calculated based on the detected bounding box. We denoted the face ratio as $ratio\_face$, and define a simple grading standard by the following formula:

$$Grade = \begin{cases} head\ shot,\ if\ (ratio\_face \geqslant 0.3) \\ close\ shot,\ if\ (0.1 \leq ratio\_face < 0.3) \\ half\ body,\ if\ (0.05 \leq ratio\_face < 0.1) \\ whole\ body,\ if\ (0.005 \leq ratio\_face < 0.05) \\ far\ shot,\ if\ (0 < ratio\_face < 0.005) \\ no\ face,\ if\ (ratio\_face = 0) \end{cases} \tag{5}$$

where the thresholds in Eq. 5 are empirically determined based on the statistical data from large amounts of single person images. Through the face-assisted classification, six subcategories are built. We regard each subcategory as a separate subdomain. The data belonging to the same subdomain are more consistent and compact than before in terms of appearance similarity, background complexity, and so on.

We then propose a multi-branch Feature Fusion Module to handle different subdomain data. As shown in Fig. 3, there are six FFM branches in total, and they are FFM_fs, FFM_wb, FFM_hb, FFM_cs, FFM_hs, and FFM_nf. Each FFM branch has the same architecture with its parameters. We train the Multi-domain TriSeNet with different subcategory data exclusively and iteratively at each mini-batch so that the network can learn different feature fusion strategies. Under different image conditions, the Multi-domain TriSeNet can choose the optimal feature fusion strategy, which improves the accuracy of our method.

C. Self-supervised Learning Strategy

To obtain the Multi-domain TriSeNet suitable for different image conditions, a straight-forward way is to train each FFM branch with different subcategory data independently. In this way, the single person segmentation task is divided into six subtasks of different subcategories. The Multi-domain TriSeNet can benefit from the multi-task learning strategy. However, these subtasks are not entirely independent of each other and the naive multi-task learning strategy ignores the domain relation among subcategory data. We propose a self-supervised learning strategy to dig out the relation among subdomains and improve the consistency of predictions through knowledge sharing.

In our self-supervised learning strategy, each FFM branch can be regarded as an expert classifier for the single person segmentation task. Taking subcategory $c \in \{1, ..., n\}$ as an example, the c^{th} FFM branch supervised by the ground-truth label Y^c is regarded as the optimal expert and other FFM branches without supervision are regarded as the candidate experts. It means that each FFM branch should not only have the power to separate the data belonging to its subdomain but also have the potential to handle other subdomain data. Letting $X^{c,l}$ be the l^{th} image of subcategory c. The soft label of $X^{c,l}$ is generated by averaging predictions of all other FFM branches. For simplicity, we omitted the superscripts of image indices. Then the generated soft label $\bar{Y^c}$ can be described as follows:

$$\bar{Y^c} = \frac{1}{n-1} \sum_{j \neq c}^{n} \sigma(P^j) \tag{6}$$

where P^j is the prediction score of the j^{th} FFM branch and σ is the Softmax operation. The multiple candidate experts can be aggregated into a stronger one, which contains multiple knowledge of other subdomains. We propose a self-supervised learning loss to improve the predictive consistency between the optimal expert and all other candidate experts:

$$L_{ssl}^{c} = -\sum_{i \in Y^c} \bar{y_i^c} \log y_i^c \tag{7}$$

where $\bar{y_i^c}$ and y_i^c are the soft label and predictions after the Softmax operation. Each branch has different training data so the self-supervised learning loss is heterogeneous for each subdomain. During the training time, the data are fed into all FFM branches, and all FFM branches can learn from both the inner-domain and cross-domain supervisions. The total loss in a mini-batch is defined as follows:

$$L = L_{seg} + \lambda_1 L_{edge} + \lambda_2 L_{ssl} \tag{8}$$

where λ_1 and λ_2 are balance factors to control the importance of the three types of losses. During the testing time, three types of feature representations are extracted by the shared paths, then the optimal FFM branch selected by Eq. 5 fuses them properly for final prediction.

V. Experiment Results

We provide the implementation details in § V-A. In § V-B, the benchmark datasets and evaluation protocols are described. The experimental results in comparison to the state-of-the-art approaches are shown in § V-C. Then we conduct ablation studies to investigate the effects of each component of our approach in § V-D. finally, several photo-editing applications benefiting from our method are shown in § V-E.

A. Implementation details

Network Architecture: We adopt a modified Xception model Xception39 and ResNet18 [7], [36] as the backbone of our method. The spatial path is composed of three 3×3 convolutions with stride 2 as CBR blocks and one 1×1 convolution as the Bottleneck block. The edge path utilizes three 1×1 convolution layers to extract multi-scale edge features and then fuses them with a linear transformation. The six FFM branches have the same network architecture but different parameters to handle single person segmentation under specific subdomains. The height and width of the final prediction are 1/8 of the input image. The bilinear interpolation operation is used to rescale the resolution.

Training: The resolution of input images is 512×512. We employ the mean subtraction, random horizontal flip, random scale, and random rotation for data augmentation during the training process. The scale rate varies from 0.9 to 1.1 and the rotate angle is between -10 and 10 degrees. The edge label is generated from the ground-truth label of the segmentation annotation by extracting the border between foreground and background areas. The soft labels for the proposed self-supervised learning strategy are generated by averaging the predictions of other FFM branches. The mini-batch stochastic gradient descent (SGD) [17] is employed and the batch size is 8. The momentum is 0.9 with weight decay $5e^{-4}$. The “poly” learning rate strategy with power 0.9 is used in training where the initial rate is multiplied by $(1 - \frac{iter}{max\_iter})^{0.9}$ at each iteration. We initialize the parameters of the whole model similar to [12]. The model is trained with the initial learning rate $2.5 \times 1e^{-2}$ for 90 epochs. We set $\lambda_1 = 1$ in Eq. 4 for the TriSeNet model and set $\lambda_1 = 1, \lambda_2 = 0.25$ in Eq. 8 for the Multi-domain TriSeNet model.

Reproducibility: The proposed method is implemented on PyTorch. All networks are tested on TESLA P100 with 16GB memory. Our method achieves 120 FPS with the ResNet18 backbone on the newly built MSSP20k dataset.

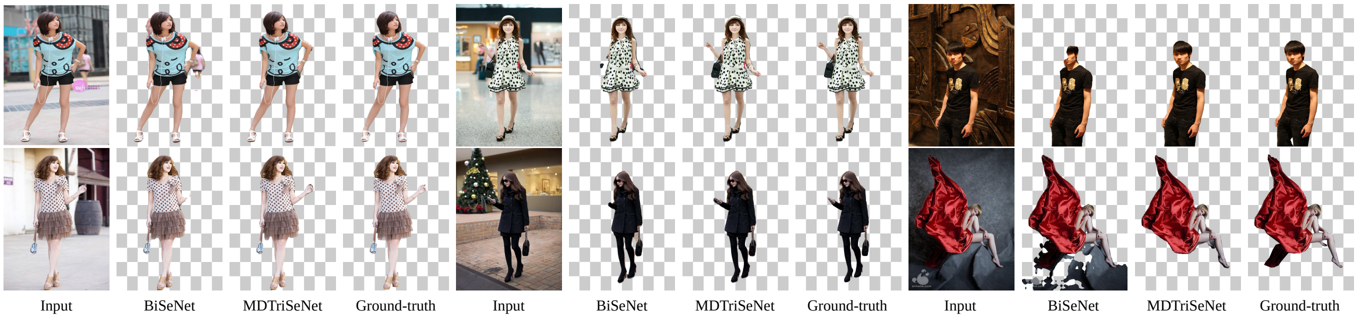


Fig. 7. Visual comparison between state-of-the-art methods [36] and our approach on the Baidu people segmentation database [33]. Our approach separates the dominant person from the background successfully when BiSeNet fails.

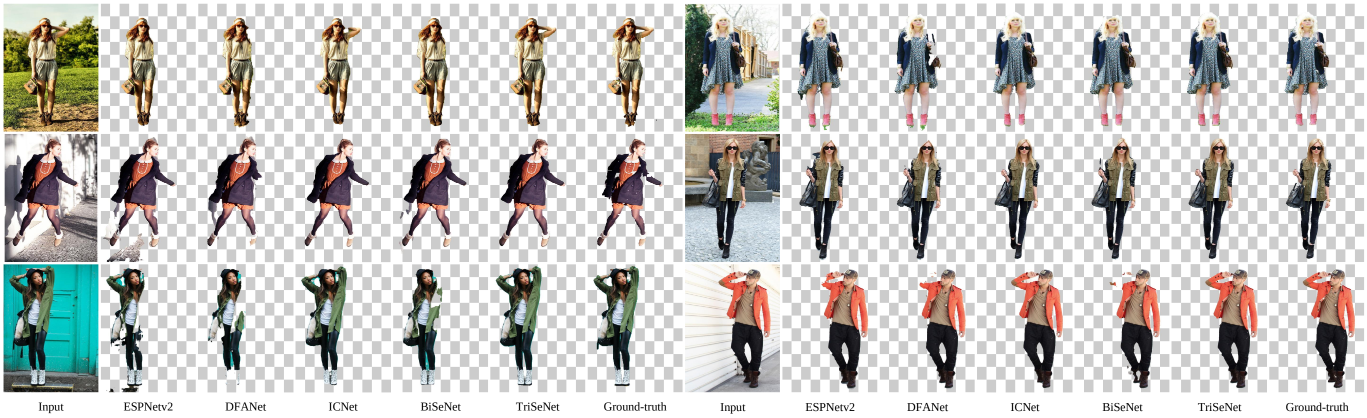


Fig. 8. Visual comparison between state-of-the-art methods [49], [50], [38], [36] and TriSeNet on the public human parsing benchmark ATR [19], [20]. Our method can separate the person with a more complete body and sharper boundaries.

B. Datasets and metrics

The overall experiments are conducted on the Baidu people segmentation database [33], the Human Parsing dataset ATR [19], and our MSSP20k dataset. Since MSSP20k is more complex than Baidu people segmentation database and ATR, we further carry out the ablation studies on it to analyze the effects of each module.

The performance of accuracy for single person segmentation is measured by Interaction-over-Union (IoU), which is the same as [33], [31]. Besides, we introduce the Contour Accuracy (CA) in DAVIS benchmarks [25], [26], [3], [4] to evaluate the accuracy of human boundaries.

The IoU calculates the overlapping rate between predictions and ground-truths. Mathematically, the score can be formulated as follows:

$$IoU = \frac{MaskPD \cap MaskGT}{MaskPD \cup MaskGT} \tag{9}$$

where $MaskPD$ and $MaskGT$ are the predicted masks and ground-truth masks of the input images.

The CA is used to measure the segmentation accuracy of the object boundaries. Given the predicted mask and ground-truth mask, the foreground contours of these two masks can be generated with the same approach as the training protocol. Then the CA is calculated by the contour-based precision P_c and recall R_c between the contour pixels:

$$CA = \frac{2P_cR_c}{P_c + R_c} \tag{10}$$

C. Comparison to the state-of-the-arts

In this section, we compare the proposed method to other state-of-the-art methods on three different datasets. IoU and CA are both utilized for accuracy comparison. We also apply FPS as the computational cost metric for speed comparison on our MSSP20k dataset.

Baidu people segmentation database. The Baidu people segmentation database consists of 5,387 training images and 1,316 test images. Followed by previous work [31], [33], 500 images from the training set are selected to form the validation set and the rest data is used for training. The Baidu people segmentation database ranks the methods according to the IoU metric. We also adopt the CA metric to evaluate the performance of boundary segmentation. We compare the proposed Multi-Domain TriSeNet (MDTriSeNet) to the other recent segmentation methods including Song *et al.* [31], Wu *et al.* [33], DFANet [50], and BiSeNet [36]. The experimental results of IoU and CA metrics are reported in Table II, where the backbone of BiSeNet and ours is ResNet18. Notice that although Wu *et al.*[33] ranked first in the Baidu people segmentation competition, our method achieves 93.49% IoU and outperforms other methods on the Baidu database. For

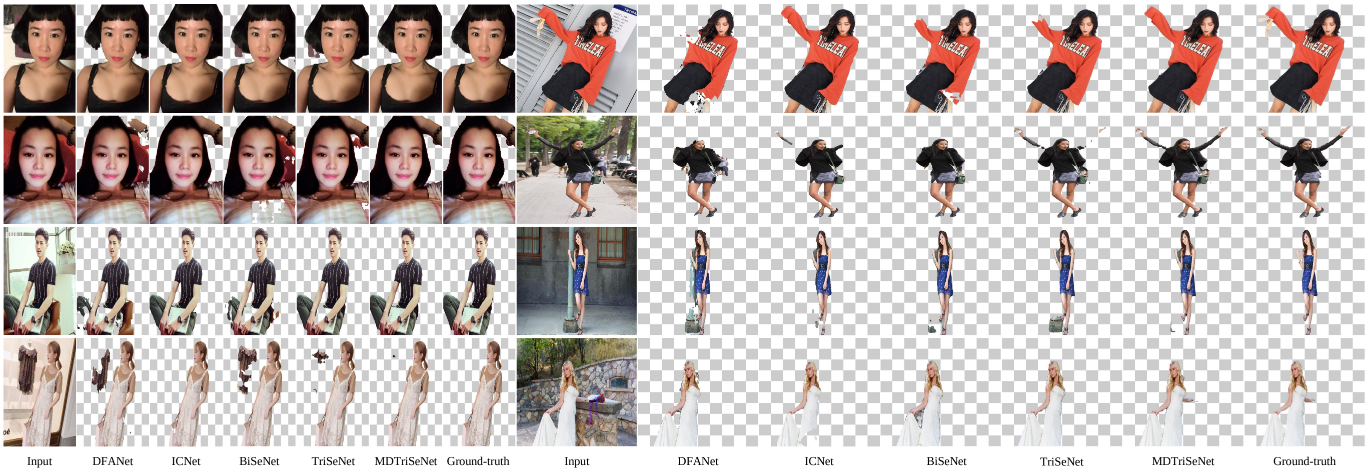


Fig. 9. Visual comparison between the state-of-the-art methods [50], [38], [36], and our approach on the MSSP20k dataset, which contains single person images with a large range of focal lengths.

TABLE II
COMPARISON RESULTS ON THE BAIDU PEOPLE SEGMENTATION DATABASE. THE BEST RESULTS ARE HIGHLIGHTED WITH **BOLD** FONT.

Method	song[31]	wu[33]	DFANet[50]	BiSeNet[36]	MDTriSeNet
IoU	83.57%	86.83%	89.36%	91.50%	**93.49%**
CA	-	-	0.701	0.734	**0.783**

boundary accuracy, our method achieves 0.783 CA. Compared to our basic model BiSeNet, the IoU score is improved from 91.50% to 93.49%, and the CA score is improved from 0.734 to 0.783, which demonstrates the effectiveness of our method.

Fig. 7 illustrates the qualitative results of the Baidu people segmentation database. Compared to the BiSeNet, our method can separate the dominant person from the background accurately with sharper human boundaries.

Human parsing benchmark ATR. ATR [19], [20] contains a total of 17, 706 single person images collecting from the Chictopia10k dataset, the Fashionista [35] dataset, and daily photos. To evaluate the performance of single person segmentation, we re-label the annotations by combining all human parts into one foreground class and the rest pixels are viewed as the background. Noting that the ATR dataset lacks headshot and far shot images, which limits the effectiveness of the proposed multi-domain learning framework. We only compare our baseline model TriSeNet to other state-of-the-art real-time methods including ESPNetv2 [49], ICNet [38], DFANet [50] and BiSeNet [36]. As shown in Table III, we select ResNet18 as the backbone of both BiSeNet and TriSeNet. BiSeNet achieves 93.71% IoU and 0.705 CA while our baseline model achieves 94.99% IoU and 0.740 CA. By introducing the edge path into BiSeNet, TriSeNet obtains 1.28% and 3.50% performance improvements in terms of IoU and CA metrics, respectively. The TriSeNet outperforms other methods in terms of the IoU metric and gets the second rank in terms of the CA metric. Notice that ICNet with the ResNet50 backbone ranks first in terms of the CA metric, but it requires more computational costs than TriSeNet.

Fig. 8 shows the visual comparison of the TriSeNet to other methods including ESPNetv2, DFANet, ICNet, and BiSeNet on the ATR benchmark. The first row shows the scenes under the strong illumination, where the light-colored hairs are hard to be segmented accurately. Some other methods fail to segment the complete human body while the TriSeNet handles the task successfully in this condition. The second row shows the scenes with cluttering background, and the last row shows the scenes with the unusual pose. Compared to other state-of-the-art unified models, our TriSeNet extracts the additional edge features for segmentation so it can generate better segmentation masks.

TABLE III
COMPARISON OF HUMAN SEGMENTATION PERFORMANCES ON THE ATR DATASET AND THE PROPOSED MSSP20K DATASET WITH OTHER STATE-OF-THE-ART METHODS INCLUDING [24], [50], [50], [50], [36], [49], [38]. THE IOU, CA AND FPS METRICS ARE APPLIED. THE BEST AND THE SECOND BEST RESULTS ARE HIGHLIGHTED WITH RED AND BLUE FONT, RESPECTIVELY.

Method	Backbone	ATR		MSSP20k		
		IoU	CA	IoU	CA	FPS
ENet[24]	-	90.15%	0.615	83.90%	0.543	158.1
DFANet[50]	XceptionA[50]	91.31%	0.692	86.52%	0.675	172.5
BiSeNet[36]	ResNet18	93.71%	0.705	86.94%	0.698	164.7
ESPNetv2[49]	-	91.82%	0.680	87.38%	0.712	81.2
ICNet[38]	ResNet50	94.61%	0.772	89.25%	0.760	56.8
TriSeNet	ResNet18	94.99%	0.740	90.43%	0.749	137.7
MDTriSeNet	ResNet18	-	-	91.52%	0.782	120.9

Our MSSP20k dataset. MSSP20k contains single person images with large variations of pose, clothing, hairstyle, focal length, and background. Table III shows the numerical comparison between our method and other state-of-the-art real-time segmentation methods including ENet [24], ESPNetv2 [49], ICNet [38], DFANet [50], and BiSeNet [36]. The IoU and CA scores are reported in Table III, together with the FPS metric for the speed analysis. The Multi-Domain TriSeNet (MDTriSeNet) achieves 91.52% IoU and 0.782 CA, which outperforms other methods. BiSeNet achieves 86.94% IoU and 0.698 CA with 164.7 FPS while TriSeNet achieves 90.43% IoU and 0.749 CA with 137.7 FPS, which improves the effectiveness of our baseline model. ICNet outperforms TriSeNet in terms of CA metric with the ResNet50 backbone.

TABLE IV
THE OVERALL COMPARISON BETWEEN THE BASIC MODEL BISENET AND OUR METHOD. GFLOPS AND PARAMETER METRICS ARE USED FOR THE SPEED ANALYSIS. THE IOU AND CA METRICS ARE USED FOR THE ACCURACY ANALYSIS.

Backbone	Method	FLOPs	Params	IoU	CA
Xception39	BiSeNet	3.53G	1.44M	85.96%	0.659
	TiSeNet	4.19G	1.57M	87.27%	0.701
	MCTriSeNet	4.19G	9.42M	87.94%	0.738
	MDTriSeNet	4.19G	2.97M	88.37%	0.742
ResNet18	BiSeNet	13.07G	12.08M	86.94%	0.698
	TiSeNet	13.76G	12.92M	90.43%	0.749
	MCTriSeNet	13.76G	77.52M	91.06%	0.775
	MDTriSeNet	13.76G	14.32M	91.52%	0.782

However, the speed of TriSeNet is two times faster than ICNet. By introducing the multi-domain learning framework and the self-supervised learning strategy, the IoU of our method is improved to 91.52%, and the CA is improved to 0.782 with a high speed of 120.9 FPS.

We further illustrate the visualization of performance on the MSSP20k dataset. The experimental results of DFANet, BiSeNet, ICNet, TriSeNet, and our full model MDTriSeNet are shown in Fig. 9, respectively. The MSSP20k dataset is composed of various single person images such as the half-body, and far-shot images. It is difficult for unified models to separate the foreground person accurately under different image subdomains. The state-of-the-art methods separating the dominant person successfully in some subdomains may generate fragmented segmentation results in other subdomains. The proposed MDTriSeNet can select the proper feature fusion module for the input image so it is more robust than other methods and can achieve better performance.

D. Ablation studies

Overall Comparison. The results of the overall comparison on the MSSP20k dataset are shown in Table IV. The Float Point Operations (FLOPs) and the scale of parameters are reported for speed analysis. The IoU and CA metrics are applied to evaluate accuracy. For a fair comparison, we set 512×512 as the resolution of input images. We choose Xception39 and ResNet18 as backbone networks. BiSeNet is the basic model of our method. TriSeNet extends the basic model by adding an edge path with the edge loss in Eq. 3. To improve the performance of TriSeNet, we classify training data into six subcategories with Eq. 5 and individually train six TriSeNet models suitable for different subcategories. We name this method as Multi-Category TriSeNet (MCTriSeNet). Specifically, MCTriSeNet consists of six TriSeNet models without parameter sharing. It can achieve higher accuracy with a six times scale of parameters than TriSeNet. To reduce the parameters of MCTriSeNet, we propose Multi-Domain TriSeNet (MDTriSeNet) as shown in Fig. 3. It utilizes the multi-branch FFM to decouple the networks into the domain-independent and the domain-specific layers so the parameters of three network paths are shared for all subcategories. Besides, the self-supervised learning strategy is designed to improve the performance of MDTriSeNet during training.

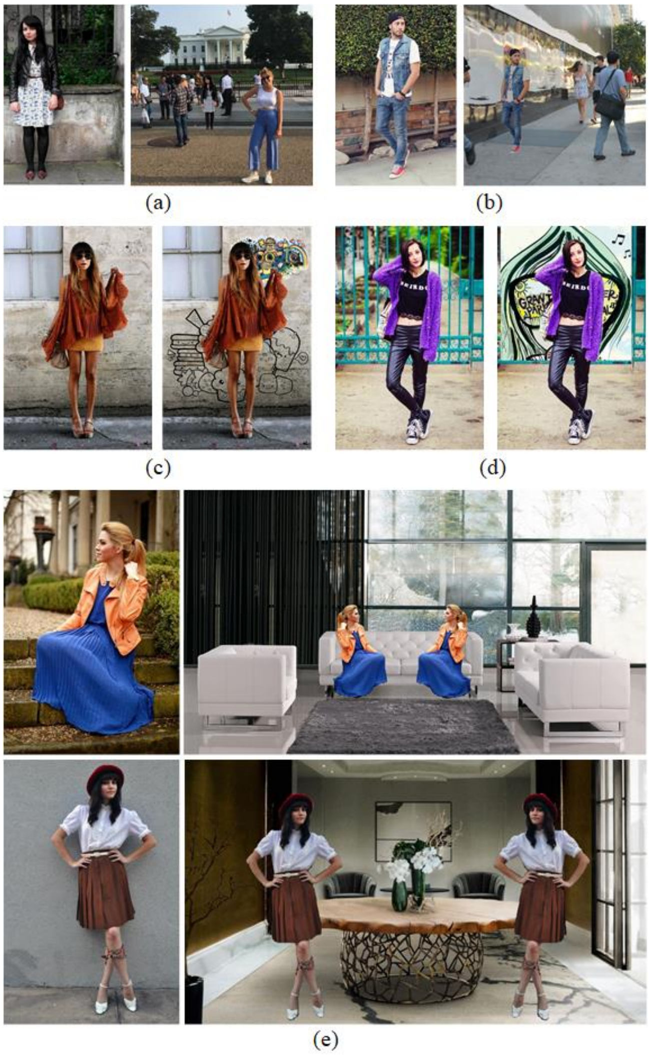


Fig. 10. Application examples of background editing (a-b), image composition (c-d), and mirror image synthesis (e). The left images stand for the input images, and the right ones are the edited results, respectively.

For the Xception39 backbone, BiSeNet achieves 85.96% IoU and 0.659 CA with 1.44M parameters. TriSeNet requires additional 0.13M parameters but obtains 1.31% IoU and 4.2% CA improvements. Compared to BiSeNet, MCTriSeNet achieves 87.94% IoU and 0.738 CA but with more than 6 times the scale of parameters. MDTriSeNet reduces the overall parameters from 9.42M to 2.97M and achieves 88.37% IoU and 0.742 CA. For the ResNet18 backbone, although BiSeNet can achieve better performance benefiting from larger backbones, our MDTriSeNet with Xception39 backbone can outperform the BiSeNet in terms of both accuracy and speed metrics. By adding an extra edge path into BiSeNet, TriSeNet obtains 3.49% IoU improvement and 5.1% CA improvements. MCTriSeNe achieves 91.06% IoU and 0.775 CA with 77.52M parameters, while the proposed MDTriSeNet achieves 91.52% IoU and 0.782 CA with only 14.32M parameters. The experimental results demonstrate that our method can be built on different backbone networks and the Multi-Domain TriSeNet is efficient to the single person segmentation task.

TABLE V
SUBCATEGORY COMPARISON OF THE PROPOSED MULTI-DOMAIN TRISENET WITH RESNET18 BACKBONE ON THE MSSP20K DATASET IN TERMS OF IOU AND CA METRICS. SSL MEANS THE PROPOSED SELF-SUPERVISED LEARNING STRATEGY. THE SUBCATEGORIES INCLUDE $far\ shot$, $whole\ body$, $half\ body$, $close\ shot$, AND $head\ shot$. THE BEST AND THE SECOND BEST RESULTS ARE HIGHLIGHTED WITH RED AND BLUE FONT, RESPECTIVELY.

Method	$far\ shot$		$whole\ body$		$half\ body$		$close\ shot$		$head\ shot$	
	IoU	CA	IoU	CA	IoU	CA	IoU	CA	IoU	CA
BiSeNet	85.64%	0.611	87.04%	0.514	84.36%	0.680	89.40%	0.746	88.48%	0.799
TriSeNet	89.78%	0.712	90.15%	0.549	89.90%	0.706	92.32%	0.820	91.01%	0.821
MCTriSeNet	90.28%	0.748	90.63%	0.584	90.84%	0.725	92.60%	0.854	91.27%	0.824
MDTriSeNet w/o SSL	90.79%	0.753	90.72%	0.583	90.77%	0.712	92.64%	0.864	91.12%	0.813
MDTriSeNet	90.98%	0.755	90.84%	0.603	91.03%	0.727	92.75%	0.858	91.67%	0.854

TABLE VI
ANALYSIS OF THE EDGE LOSS WITH A DIFFERENT BALANCING WEIGHT IN EQ. 4. $\lambda_1 = 0$ MEANS THE MODEL IS TRAINED WITHOUT THE EDGE LOSS. THE BEST RESULTS ARE HIGHLIGHTED WITH **BOLD** FONT.

	$\lambda_1 = 1$	$\lambda_1 = 0.5$	$\lambda_1 = 0.25$	$\lambda_1 = 0$
IoU	**90.43%**	90.33%	89.88%	87.57%
CA	**0.749**	0.738	0.730	0.714

TABLE VII
ANALYSIS OF THE MXUNIT OF THE SPATIAL PATH. THE BEST RESULTS IN EACH METHOD ARE HIGHTED WITH **BOLD** FONT.

Method	ReLU	mxUnit	IoU	CA
TiSeNet	✓		90.12%	0.725
		✓	**90.43%**	**0.749**
MDTriSeNet	✓		91.16%	0.761
		✓	**91.52%**	**0.782**

Analysis of edge loss. During the training process, the edge loss is utilized to help edge path extract boundary information of the body. We evaluate the effectiveness of the loss function on our baseline model TriSeNet and report the IoU and CA metrics in Table VI. We select ResNet18 as the backbone networks. λ_1 is the weight of edge loss in Eq. 4. $\lambda_1 = 0$ means the condition that the model is trained without edge loss. When TriSeNet is trained without the edge loss, the improvement of accuracy is limited. With the help of edge loss, the IoU is improved from 87.57% to 90.43%. In addition, the CA is improved from 0.714 to 0.749. For a fair comparison, we finally fix $\lambda_1 = 1$ for both MCTriSeNet and MDTriSeNet models in the other experiments.

Analysis of mxUnit. The mxUnit is proposed to generate the discriminative feature representation for the spatial path. We compare the effectiveness of mxUnit with the ReLU activation, and report the results in Table VII. The backbone networks of the TriSeNet and the MDTriSeNet are ResNet18. Compared to the ReLU activation, the mxUnit aggregates locally spatial information and brings about 0.31% IoU and 2.4% CA improvement for TriSeNet, and 0.36% IoU and 2.1% CA improvement for MDTriSeNet, which demonstrates the effectiveness of the mxUnit.

Subcategory Comparison. The multi-domain learning framework and the self-supervised learning strategy are proposed to improve the accuracy of our method under different image conditions. As described in Section IV-B, single person images are classified into six subcategories. Each subcategory can be regarded as a separate image subdomain. We evaluate our method on these subdomains and report the results in Table V. Compared to BiSeNet, TriSeNet obtains significant performance improvement under all subdomains. In particular, the $half\ body$ subdomain obtains 5.54% IoU and 2.6% CA improvement, which ranks the first among all subdomains. The results of TriSeNet prove that the edge information of the human body is essential to the single person segmentation task. MCTriSeNet has six times of parameters than TriSeNet. The networks are trained with different subcategory data individually so it can achieve better performance than TriSeNet by learning more fine-grained feature representations. MDTriSeNet shares the spatial path, the context path, and the edge path for all subcategories, and all FFM branches are trained jointly. The parameters of these three paths are trained through the gradient from all training data, which contributes to learning both the domain-independent and the domain-specific representations. Even without the proposed Self-Supervised Learning (SSL) loss, MDTriSeNet can achieve a comparable performance but fewer parameters in comparison to MCTriSeNet. In detail, MDTriSeNet without SSL outperforms MCTriSeNet in $far\ shot$, $whole\ body$, and $close\ shot$ subdomains in terms of IoU. By applying the self-supervised learning loss, each FFM branch is not only trained with subdomain data directly but also able to learn knowledge from other branches, which improves the predictive consistency among all branches. After introducing the self-supervised learning loss during training, the performance of MDTriSeNet is improved in all subdomains in terms of the IoU metric.

E. Human-based Applications

With the proposed segmentation method, users can easily separate the foreground from the background in single person images. In this section, we demonstrate some examples of background editing and image synthesis. As shown in Fig. 10, since the segmentation results of our method are quite accurate, the results of photo editing are visually harmonious. The first row shows the example of background editing, where the dominant person is segmented via the proposed method and moved to a new scene. The second row shows the example of image composition. The foreground person is separated from the background and extra elements are added to the background. Finally, the foreground is restored and more appealing components are shown in the new image. The last two rows are the example of comprehensive application such as background replacement, foreground rescaling, and object duplication.

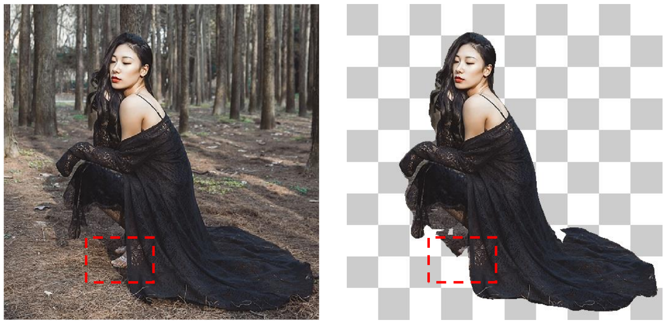


Fig. 11. The failure case of the proposed method. The inaccurate area is highlighted in the red dashed bounding box.

The proposed method aims to tackle the single person segmentation task by aggregating multiple cues and mining the relation among different subdomains. However, the proposed method may face failure for some challenging scenes unrelated to these motivations. For example, as shown in Fig. 11, a segmentation failure occurs due to the lack of discrimination in shadow areas. To avoid such failure cases, the cue about the shadow detection may be explored in the future.

VI. Conclusion

In this paper, we have proposed a real-time single person image segmentation method with high performance. The novel baseline TriSeNet consists of three network paths to extract the high-dimensional spatial features, high-level semantic features, and detailed boundary features jointly. TriSeNet focuses on both the foreground content and the human edges, so it can generate a high-quality foreground mask with sharp boundaries. We then present a multi-domain learning framework to construct the Multi-domain TriSeNet. By introducing the multi-branch Feature Fusion Module (FFM), both the domain-independent and the domain-specific representations can be learned during training. A self-supervised learning strategy is proposed to further enhance performance by improving predictive consistency among FFM branches. Moreover, we collect and build a new dataset MSSP20k for the real-world single person segmentation task. The performance on the public and the new datasets demonstrate that our Multi-domain TriSeNet outperforms the state-of-the-art methods with high speed.

Acknowledgments

Xiaogang Jin was supported by the National Natural Science Foundation of China (Grant Nos. 62036010, 61972344), the Key Research and Development Program of Zhejiang Province (Grant No. 2018C03055), and the Ningbo Major Special Projects of the "Science and Technology Innovation 2025" (Grant No. 2020Z007). Zhiyuan Liang and Xiaogang Jin were supported by Alibaba-Zhejiang University Joint Institute of Frontier Technologies. This work was supported by Alibaba Group through Alibaba Innovative Research Program.

References

[1] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, 2017.

[2] J. Shen, J. Peng, L. Shao, "Submodular trajectories for better motion segmentation in videos," *IEEE Trans. Image Process.*, vol. 27, no. 6, pp. 2688-2700, 2018.

[3] S. Caelles, A. Montes, K.-K. Maninis, Y. Chen, L. Van Gool, F. Perazzi, and J. Pont-Tuset, "The 2018 davis challenge on video object segmentation," *arXiv:1803.00557*, 2018.

[4] S. Caelles, J. Pont-Tuset, F. Perazzi, A. Montes, K.-K. Maninis, and L. Van Gool, "The 2019 davis challenge on vos: Unsupervised multi-object segmentation," *arXiv:1905.00737*, 2019.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, 2018.

[6] X. Chen, D. Qi, and J. Shen, "Boundary-aware network for fast and high-accuracy portrait segmentation," *arXiv preprint arXiv:1901.03814*, 2019.

[7] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *IEEE CVPR*, 2017, pp. 1251–1258.

[8] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "the cityscapes dataset for semantic urban scene understanding," in *IEEE CVPR*, 2016, pp. 3213–3223.

[9] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge: A retrospective," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2015.

[10] C. Gan, M. Lin, Y. Yang, G. D. Melo, and A. G. Hauptmann, "Concepts not alone: exploring pairwise relationships for zero-shot video activity recognition," in *AAAI*, vol. 30, no. 1, 2016.

[11] K. Gong, X. Liang, D. Zhang, X. Shen, and L. Lin, "Look into person: Self-supervised structure-sensitive learning and a new benchmark for human parsing," in *IEEE CVPR*, 2017, pp. 932–940.

[12] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *IEEE ICCV*, 2015, pp. 1026–1034.

[13] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "Mobilenets: Efficient convolutional neural networks for mobile vision applications," *arXiv preprint arXiv:1704.04861*, 2017.

[14] Y. Jun Koh and C.-S. Kim, "Cdts: Collaborative detection, tracking, and segmentation for online multiple object segmentation in videos," in *IEEE ICCV*, 2017, pp. 3601–3609.

[15] I. Kligvasser and T. Michaeli, "Dense xunit networks," *arXiv preprint arXiv:1811.11051*, 2018.

[16] I. Kligvasser, T. Rott Shaham, and T. Michaeli, "xunit: Learning a spatial activation function for efficient image restoration," in *IEEE CVPR*, 2018, pp. 2433–2442.

[17] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012, pp. 1097–1105.

[18] X. Li, Z. Liu, P. Luo, C. Change Loy, and X. Tang, "Not all pixels are equal: Difficulty-aware semantic segmentation via deep layer cascade," in *IEEE CVPR*, 2017, pp. 3193–3202.

[19] X. Liang, S. Liu, X. Shen, J. Yang, L. Liu, J. Dong, L. Lin, and S. Yan, "Deep human parsing with active template regression," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 12, pp.2402–2414, 2015.

[20] X. Liang, C. Xu, X. Shen, J. Yang, S. Liu, J. Tang, L. Lin, and S. Yan, "Human parsing with contextualized convolutional neural network," in *IEEE ICCV*, 2015, pp. 1386–1394.

[21] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *ECCV*, 2014, pp. 740–755.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *IEEE CVPR*, 2015, pp. 3431–3440.

[23] J. Lv, X. Shao, J. Xing, P. Liu, X. Zhou, and X. Zhou, "Hierarchical bilinear network for high performance face detection," in *IEEE ICIP*, 2017, pp. 415–419.

[24] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "Enet: A deep neural network architecture for real-time semantic segmentation," *arXiv preprint arXiv:1606.02147*, 2016.

[25] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *IEEE CVPR*, 2016, pp. 724–732.

[26] J. Pont-Tuset, F. Perazzi, S. Caelles, P. Arbeláez, A. Sorkine-Hornung, and L. Van Gool, "The 2017 davis challenge on video object segmentation," *arXiv:1704.00675*, 2017.

[27] L. Qi, J. Huo, L. Wang, Y. Shi, and Y. Gao, "A mask based deep ranking neural network for person retrieval," in *IEEE ICME*, 2019, pp. 496–501.

[28] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, 2015, pp. 234–241.

[29] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE CVPR*, 2018, pp. 4510–4520.

[30] X. Shen, A. Hertzmann, J. Jia, S. Paris, B. Price, E. Shechtman, and I. Sachs, "Automatic portrait segmentation for image stylization," *Comput. Graph. Forum*, vol. 35, no. 2, pp. 93–102, 2016.

[31] C. Song, Y. Huang, Z. Wang, and L. Wang, "1000fps human segmentation with deep convolutional neural networks," in *ACPR*, 2015, pp. 474–478.

[32] P. Voigtlaender, M. Krause, A. Osep, J. Luiten, and B. Leibe, "Mots: Multi-object tracking and segmentation," in *IEEE CVPR*, 2019, pp. 7942–7951.

[33] Z. Wu, Y. Huang, Y. Yu, L. Wang, and T. Tan, "Early hierarchical contexts learned by convolutional networks for image segmentation," in *ICPR*, 2014, pp. 1538–1543.

[34] Z. Wu, C. Shen, and A. v. d. Hengel, "Real-time semantic image segmentation via spatial sparsity," *arXiv preprint arXiv:1712.00213*, 2017.

[35] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *IEEE CVPR*, 2012, pp. 3570–3577.

[36] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *ECCV*, 2018, pp. 325–341.

[37] S.-H. Zhang, X. Dong, H. Li, R. Li, and Y.-L. Yang, "Portraitnet: Real-time portrait segmentation network for mobile device," *Comput. Graph.*, vol. 80, pp. 104–113, 2019.

[38] H. Zhao, X. Qi, X. Shen, J. Shi, and J. Jia, "Icnet for real-time semantic segmentation on high-resolution images," in *ECCV*, 2018, pp. 405–420.

[39] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *IEEE CVPR*, 2013, pp. 3586–3593.

[40] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, and P. H. Torr, "Conditional random fields as recurrent neural networks," in *IEEE ICCV*, 2015, pp. 1529–1537.

[41] Q. Chen, T. Ge, Y. Xu, Z. Zhang, X. Yang, and K. Gai, "Semantic human matting," in *ACM Multimedia*, 2018, pp. 618–626.

[42] F. Pan, I. Shin, F. Rameau, S. Lee, and I. S. Kweon, "Unsupervised intra-domain adaptation for semantic segmentation through self-supervision," in *IEEE CVPR*, 2020, pp. 3763–3772.

[43] B. Zhu, Y. Chen, J. Wang, S. Liu, B. Zhang, and M. Tang, "Fast deep matting for portrait animation on mobile phone," in *ACM Multimedia*, 2017, pp. 297–305.

[44] H. Nam and B. Han, "Learning multi-domain convolutional neural networks for visual tracking," in *IEEE CVPR*, 2016, pp. 4293–4302.

[45] J. Hoffman, B. Kulis, T. Darrell, and K. Saenko, "Discovering latent domains for multisource domain adaptation," in *ECCV*, 2012, pp. 702–715.

[46] A. S. Sebag, L. Heinrich, M. Schoenauer, M. Sebag, L. F. Wu, and S. J. Altschuler, "Multi-domain adversarial learning," in *ICLR*, 2019.

[47] Y. Yang and T. M. Hospedales, "A unified perspective on multi-domain and multi-task learning," in *ICLR*, 2015.

[48] T. Xiao, H. Li, W. Ouyang, and X. Wang, "Learning deep feature representations with domain guided dropout for person re-identification," in *IEEE CVPR*, 2016, pp. 1249–1258.

[49] S. Mehta, M. Rastegari, L. G. Shapiro, and H. Hajishirz, "Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network," in *IEEE CVPR*, 2019, pp. 9190–9200.

[50] H. Li, P. Xiong, H. Fan, and J. Sun, "Dfanet: Deep feature aggregation for real-time semantic segmentation," in *IEEE CVPR*, 2019, pp. 9522–9531.

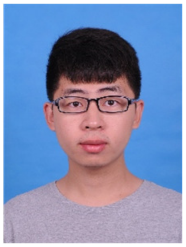
[51] J. Miao, Y. Wei, and Y. Yang, "Memory aggregation networks for efficient interactive video object segmentation," in *IEEE CVPR*, 2020, pp. 10366–10375.

[52] J. Miao, Y. Wei, Y. Wu, C. Liang, G. Li, and Y. Yang, "Vspw: A large-scale dataset for video scene parsing in the wild," in *IEEE CVPR*, 2021, pp. 4133-4143.

[53] R. Quan, L. Zhu, Y. Wu, and Y. Yang, "Holistic lstm for pedestrian trajectory prediction," *IEEE Trans. Image Process.*, vol. 30, pp. 3229–3239, 2021.

[54] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *IEEE CVPR*, 2020, pp. 9141–9150.

[55] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *IEEE CVPR*, 2018, pp. 7132–7141.



Zhiyuan Liang is working toward the Ph. D. degree in the School of Computer Science, Beijing Institute of Technology, Beijing, China. He is currently internshipping in Alibaba Group. His current research interests include video segmentation and object tracking.

Kan Guo received the Ph. D. degree from Beihang University in 2018. He is currently a senior algorithm engineer in Alibaba Group, Hangzhou, China. His current research interests include image segmentation, machine learning, and artificial intelligence.

Xiaobo Li received the B.E. degree from PLA Information Engineering University in 2000, and M.E. degree in computer science from Peking University in 2009. He joined Alibaba Group, Hangzhou, China, in 2009. He is currently a Staff Algorithm Engineer and is leading the multimedia algorithm team in Alibaba Taobao (China), Co., Ltd. His research interests include compute vision, artificial intelligence, and high efficiency video coding.

Xiaogang Jin (M'04) is a Professor in the State Key Laboratory of CAD&CG, Zhejiang University. He received the B.Sc. degree in computer science and the M.Sc. and Ph.D degrees in applied mathematics from Zhejiang University, P. R. China, in 1989, 1992, and 1995, respectively. His current research interests include image processing, traffic simulation, collective behavior simulation, cloth animation, virtual try-on, digital face, implicit surface modeling and applications, creative modeling, computer-generated marbling, sketch-based modeling, and virtual reality. He received an ACM Recognition of Service Award in 2015 and two Best Paper Awards from CASA 2017 and CASA 2018. He is a member of the IEEE and ACM.

Jianbing Shen (M'11-SM'12) is currently acting as the Lead Scientist at the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. He is also an adjunct Professor with the School of Computer Science, Beijing Institute of Technology, Beijing, China. He received his Ph.D. from the Department of Computer Science, Zhejiang University in 2007. He has published more than 100 top journal and conference papers, and ten papers are selected as the ESI Hightly Cited or ESI Hot Papers. His current research interests are in the areas of deep learning for video analysis, computer vision for autonomous driving, deep reinforcement learning, and machine learning for intelligent systems.

He is a Senior Member of IEEE. He obtained many flagship honors including the Fok Ying Tung Education Foundation from Ministry of Education, the Program for Beijing Excellent Youth Talents from Beijing Municipal Education Commission, and the Program for New Century Excellent Talents from Ministry of Education. He is an Associate Editor of *IEEE Transactions on Image Processing*, *IEEE Transactions on Neural Networks and Learning Systems*, *Pattern Recognition*, and other journals.