IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 29, 2021, accepted October 27, 2021, date of publication November 4, 2021,
date of current version December 2, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3125356

Cascaded Segmented Matting Network for Human Matting

BO LIU[1,2], (Senior Member, IEEE), HAIPENG JING[1], GUANGZHI QU[3], (Senior Member, IEEE), AND HANS W. GUESGEN[2]

[1]Faculty of Information Technology, School of Software Engineering, Beijing University of Technology, Beijing 100124, China
[2]School of Fundamental Sciences, Massey University, Palmerston North 4474, New Zealand
[3]Computer Science and Engineering Department, Oakland University, Rochester, MI 48309, USA

Corresponding author: Bo Liu (liubo03@gmail.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 62076015, and in part by the REaDI Fund from Massey University.

ABSTRACT Human matting, high quality extraction of humans from natural images, is crucial for a wide variety of applications such as virtual reality, augmented reality, entertainment and so on. Since the matting problem is an ill-posed problem, most previous methods rely on extra user inputs such as trimap or scribbles as guidance to estimate alpha value for the pixels that are in the unknown region of the trimap. This phenomenon makes it difficult to be applied to large scale data. In order to solve these problems, we studied the unique role of semantics and details in image matting, and decomposed the matting task into two sub-tasks: trimap segmentation based on high-level semantic information and alpha regression based on low-level detailed information. Specifically, we proposed a novel Cascaded Segmented Matting Network (CSMNet), which uses a shared encoder and two separate decoders to learn these two tasks in a collaborative way to achieve the end-to-end human image matting. In addition, we established a large-scale dataset with 14,000 fine-labeled human matting images. A background dataset is also built to simulate real pictures. Comprehensive empirical studies on above datasets demonstrate that CSMNet could produce a stable and accurate alpha matte without the input of trimap and achieve an evaluation value that is comparable to the algorithm that requires trimap.

INDEX TERMS Human matting, semantic segmentation, salient object detection.

I. INTRODUCTION

Human matting refers to accurately extracting the humans from natural images. It has a wide range of applications in image processing and editing, as well as in real life. The human matting could be conducted with the help of image segmentation, but the effect is not satisfied. For example, as seen in Fig.1, we need to replace the background of the girl's photo with a pure white background. As we know, image segmentation algorithm predicts a discrete semantic label for each pixel. Although the girl can be distinguished from the background, hard segmentation results lose the details of the boundary such as hair, and the final composite image effect will have a rigid edge transition. However, the soft segmentation results obtained with the help of human matting will predict a floating-point type transparency value for each pixel. The soft segmentation results contain details of the border such as hair, and the synthesis effect is smoother.

For image matting, an image I is assumed to be a linear combination of foreground F and background B via a soft alpha matte α [1].

$$I_z = \alpha_z F_z + (1 - \alpha_z) B_z \tag{1}$$

α_z represents the weight of the linear combination of the foreground color and the background color at the pixel z in the range of [0,1]. Its physical meaning is the degree of opacity of the foreground object corresponding to the pixel z. The alpha value is 1 and 0 of which pixels located in foreground region and background region respectively. For the unknown area, the alpha value is a floating decimal between 0 and 1.

The associate editor coordinating the review of this manuscript and approving it for publication was Mehul S Raval.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

FIGURE 1. Comparison of hard segmentation and soft segmentation composite image.

It means that the pixel corresponding to the foreground object is semitransparent. The goal of the human matting is to assign each pixel a more fine-grained float opacity value of foreground.

According to the mathematical model, there are 7 unknown variables but only 3 known variables for RGB image, and thus, this decomposition is severely under constrained [2]. Specifically, the 7 unknown variables include F_z and B_z for each channel and the same value α_z for the image. And the 3 known variables are the different I_z value for each channel. Therefore, most matting algorithms [3]–[6] need to take user designated trimaps or scribbles as extra constraints,wherein the user segments the input image into three regions: definite foreground, definite background, and unknown region. This three-level map is called a trimap [7]. However, the trimap is costly for humans to annotate, or suffer from low precision if captured via a depth camera.

With the development of deep learning, many researchers employ semantic segmentation to generate trimap automatically and feed the segmentation results as trimaps into the matting network. However, it is found through experiments that a high-precision trimap segmentation map cannot be obtained based on a single RGB input and a single segmentation model. According to the literature [17], high- precision trimap is conducive to obtaining more accurate alpha matte, vice versa. Therefore, how to obtain a higher precision trimap is a problem that needs to be solved at present. As far as we know, portraits will be in a significant part of portrait selfies. The three regions in trimap have semantic gaps in portrait selfies. When a segmentation network is used directly for classification, the segmentation of unknown regions is bad. Therefore, we came up with an idea to first employ saliency detection to obtain the absolute foreground and absolute background binary mask. Then use high-precision trimap labels to further classify the unknown area.

In the process of human matting in an end-to-end manner, a direct method is to train the two tasks of trimap segmentation and alpha regression separately, and input the trimap result into the alpha regression network as auxiliary information. However, this intuitive method does not work well. The reason is that the purpose of semantic segmentation is to classify each pixel and be able to roughly distinguish people from the background, while the alpha regression task is not. In response to this problem, this paper proposes a novel Cascaded Segmented Matting Network (CSMNet), which uses a shared encoder and two separate decoders to learn these two tasks in a collaborative way to achieve the end-to-end human image matting. At the same time, in the alpha regression task, the Holistic Attention Module (HAM) is added to extract the unknown area more accurately, and refine branch is introduced to generate higher precision alpha matte.

The main contributions of this paper are summarized in the following three aspects:

- To the best of our knowledge, CSMNet is the first trimap segmentation algorithm based on saliency detection technology and solved the trimap segmentation problem step by step. Empirical studies show that the trimap generated by the algorithm has high accuracy, and is better than the single DeepLabV3+ [12] network with coarse trimap result. The mIoU is increased by 1.27%.
- As a novel automatic matting structure, CSMNet is the first cascaded network structure designed to jointly solve trimap segmentation and alpha regression tasks. Under the condition of single RGB input, high-precision alpha matte and excellent hair fine effects can be obtained. The Mean Squared Error (MSE) and Sum of Absolute Differences(SAD) values reach 0.0118 and 0.3252, respectively.
- A large-scale human matting dataset is constructed. It contains 13,710 unique human images with corresponding alpha mattes. In order to simulate real data, a large background data set was constructed at the same time, covering scenes such as cars, indoors, and streets, with a total of 32,353 pictures. The dataset contributes to the development of human matting.

II. RELATED WORK

A. TRIMAP SEGMENTATION

The problem of obtaining trimap hinders the intelligent progress of image matting. The automatic generation of trimap has now become an important part of the image matting. The traditional trimap generation method mainly relies on special equipment for automatical generation, for example, the depth information of a given picture is obtained with the help of a depth-of-field camera to estimate the trimap [23]. Cameras with different shooting angles are used to shoot objects in a fixed scene at a fixed time. Then the trimap is generated according to the image changes [24]. This type of technology requires the support of expensive equipment, so it's hard to be widely used. In recent years, the development of deep learning has also extended to the field of automatically generating trimap. Semantic segmentation can provide category classification for each pixel of a given image. Naturally, trimap could be regarded as a three-category segmentation task. Meanwhile, the semantic segmentation task can meet the requirements of real-time output of trimap three-part graph, so many trimap segmentation methods based on deep learning have emerged [7], [8].

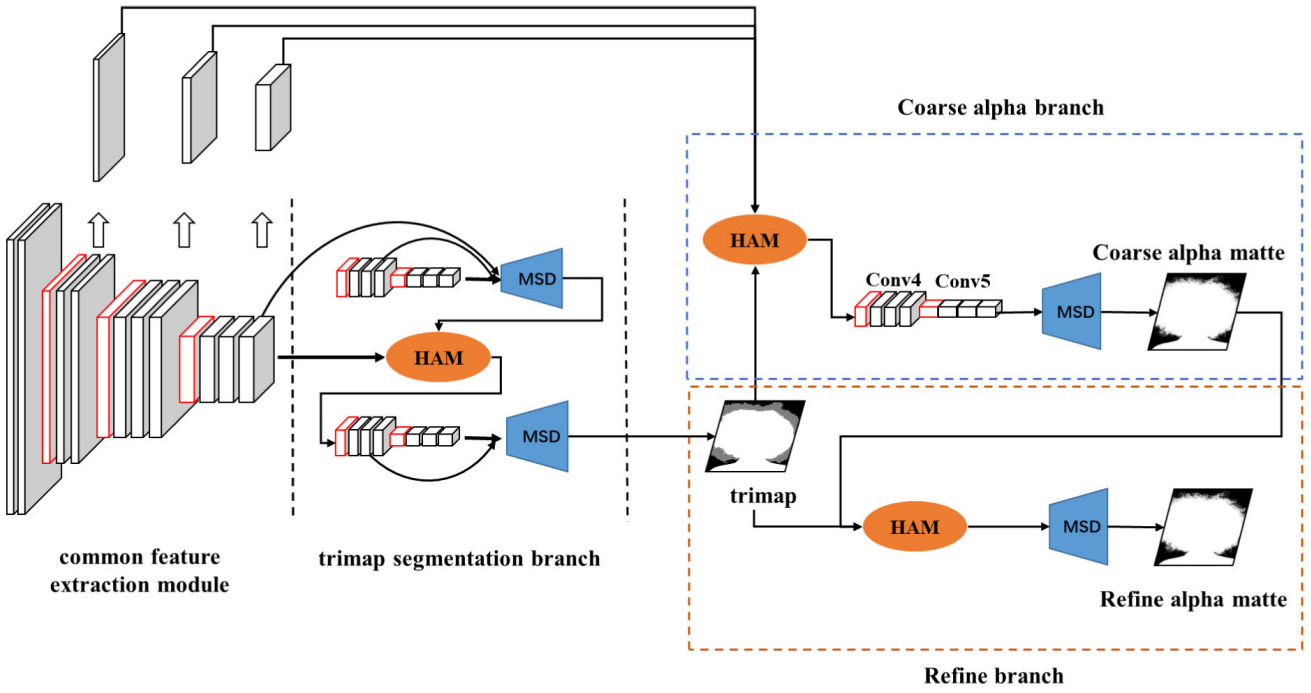

FIGURE 2. A schematic diagram of CSMNet. It can be seen that there are mainly four parts: common feature extraction module, trimap segmentation branch,coarse alpha branch and refine alpha branch.

B. SEMANTIC SEGMENTATION

Image segmentation is closely related to image matting technology. The following describes some of the research progress of image segmentation in recent years. The Full Convolutional Network (FCN) proposed by Long *et al.* [9] transfers semantic segmentation to an end-to-end manner, achieves dense prediction in pixel level and greatly improves segmentation accuracy. After that, FCN became the main framework of semantic segmentation. PSPNet [10] introduced the pyramid pool module in the full convolutional segmentation network, and obtained feature map with different resolutions by performing multiple levels of pooling operations, so as to better obtain global and contextual prior information. Peng *et al.* [11] pointed out that the use of large convolution kernels and boundary optimization modules could improve the classification accuracy at the pixel level while maintaining accurate positioning capabilities. In 2018, Chen *et al.* released DeepLabV3+ [12], which employed an encoder-decoder architecture, and introduced a deep separable convolution operation. The best DeepLabv3+ got 89.0% mIoU score in COCO and 2012 PASCAL VOC challenge.

C. IMAGE MATTING

Traditional matting algorithms heavily rely on low-level features, e.g., color cues, to determine the alpha matte through sampling [3], [4], [13] or propagation [6], which often fail in complex scenes.

Deep Image Matting (DIM) [14] pioneered the application of deep learning to solve the task of image matting. Meanwhile, DIM also created a large-scale matting data set. This model combines RGB images with trimap as input, employs high-level semantics to estimate the matting alpha matte. [15] proposed a hybrid sampling and learning-based image matting method. Other studies [16], [17] established two branches to perceive alpha matte, and these two branches enhanced each other to perfect the final result. However, all of these image matting networks rely on trimap to enhance their semantic refinement. Since obtaining a trimap requires user effort, some recent methods (including our CSMNet) attempt to avoid it, as described below.

Image matting is extremely difficult when trimaps are unavailable since semantic estimation will be necessary (to locate the foreground) before predicting a precise alpha matte [20]. Currently, trimap-free methods always focus on a specific type of foreground objects, such as humans. Some matting methods [7], [8] employ image segmentation to generate trimap, and design networks to complete trimap segmentation tasks and matting tasks. But the generation of trimap will be of poor quality which affects the performance of the subsequent matting branch. Yang, *et al.* [21] uses LSTM and reinforcement learning to generate trimap, but it requires simple user interaction and more response time. Zhang, *et al.* [22] counted two branches to extract the specific information of the foreground and background respectively, and designed a fusion module to fuse the output information of the front background to generate an alpha mask. Although they realize the automatic matting without adding additional auxiliary information, how to better implement the trimap segmentation map has not been designed and

discussed in detail, because the trimap auxiliary information is essential for obtaining high-precision matting masks.

III. METHODS

This section mainly introduces the philosophy and network structure of CSMNet. First, a trimap segmentation method based on saliency detection was proposed. Then, each part of CSMNet was introduced in detail.

A. OVERVIEW OF THE PROPOSED MODEL

At present, most of the mainstream trimap segmentation algorithms are assisted by semantic segmentation and the output of the fully convolutional neural network as the trimap directly. However, from an experimental point of view, the best segmentation network is DeepLabV3+ with only RGB images input. However, the segmentation output is not satisfactory, and it is difficult to obtain a high-precision trimap. In order to verify the deficiencies of the semantic segmentation in the trimap segmentation algorithm and elicit our solution, we first conducted a verification test which can be seen in Section IV.A.

This paper proposed a cascaded trimap segmentation network based on saliency detection which can be decomposed into three parts: 1) the common feature extraction network; 2) the trimap segmentation branch; 3) the alpha regression branch which includes coarse alpha branch and refine branch.

The schematic diagram of the model is shown in Fig.2. The first part of the common feature extraction network provides public shallow feature information for the remaining two parts, so as to achieve the greatest degree of feature reuse. The remaining two parts perform their duties, and obtain the saliency detection mask map and the trimap segmentation map respectively. The three parts progress step by step, and the output of the previous part is used as the input of the next part to cooperate to complete the trimap segmentation task.

B. CASCADED SEGMENTED MATTING NETWORK

This section will introduce each part of the proposed CSMNet in detail.

1) COMMON FEATURE EXTRACTION MODULE

In the common feature extraction branch, VGG16 [25] network is employed in our model. The purpose of the branch is to extract shallow, common feature information for later usage. The common feature extraction network selects the first three convolutional layers of the VGG-16 network as component modules, including Conv1, Conv2, and Conv3. After these three convolutional layers, feature maps with 64, 128, and 256 channels will be obtained respectively. The resolution of the feature map is the input size, 1/2 of the input size, and 1/4 of the input size, which can be expressed as $\{f_i, i = 1, \ldots, 3\}$.

2) TRIMAP SEGMENTATION BRANCH

When an input image I is given, the public feature extraction network g_c first obtains the shallow feature information map S, that is, $g_c(\mathrm{I})$. Then the saliency detection branch g_s generates a probability map P with the same resolution as S. The higher the probability, the more likely the pixel is to be a salient object, and finally the saliency mask map S_m is obtained. Finally, the shallow feature information map S and the saliency mask map S_m are fused into the trimap prediction branch network g_t to obtain the final trimap segmentation map T, which can be expressed by the following formula:

$$\mathrm{S} = \mathrm{g_c(I)} \tag{2}$$

$$\mathrm{S_m} = \mathrm{g_s(I)} \tag{3}$$

$$T = g_t\left[HAM\left(S, S_m\right)\right] \tag{4}$$

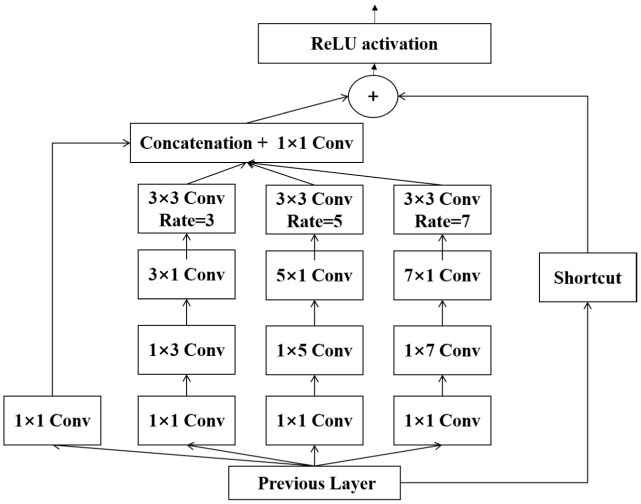

FIGURE 3. A schematic diagram of SPRFB module.

a: MULTI-SCALE FUSION DECODER (MSD)

For pixel-level classification tasks, multi-scale information fusion plays a vital role in improving the final effect [26]. Multi-scale information integrates feature information at different sizes, reorganizes global semantic information and detailed feature information, and is widely used in image segmentation and saliency detection. Inspired by Receptive Field Block (RFB) [27], highly aggregated lightweight module named Special RFB (SPRFB) is proposed to strengthen the feature extraction ability of the network. Meanwhile, in order to make full use of multi-level feature information, a feature aggregation module is also designed, which will be introduced in turn below.

- SPRFB needs to process the feature information in multiple levels, so this structure first reduces the dimensionality of the level information to reduce model complexity.
- The RFB module adopts standard convolution operations, and SPRFB adopts depthwise separable convolution, whose parameters reduced by 1/9 compared with the standard convolution operation.
- Compared with the RFB module, SPRFB module adds a branch to only perform 1×1 convolution, in order to further increase the receptive field.

- It introduced the residual structure design commonly used in Inception-ResNet V2 [28] and ResNet [29], which effectively suppresses the problem of vanishing/exploding gradients.

The structure of the SPRFB module is shown in Fig.3.

After the SPRFB module, multiple feature maps can be obtained. In order to output the feature map with the specified number of channels, we adopt a strategy of multiplying different levels in pixelwise for feature aggregation.

For the low-level feature map f_{3d}, we first perform up-sampling with the corresponding size, and pass the 3×3 conv operation, and then multiply it with the middle-level f_{4d} feature in an element-wise manner to obtain f_{4d_1} feature map. At the same time, f_{3d} goes through up-sampling of factor 4 with 3×3 conv operation, and f_{4d_1} undergoes up-sampling of factor 2 with 3×3 conv operation. Finally, the above all multiply with feature f_{5d} in an element-wise manner to obtain f_{5d_1}. With the same principle, we finally get a feature map fused with multi-layer features named f_{5d_2}.

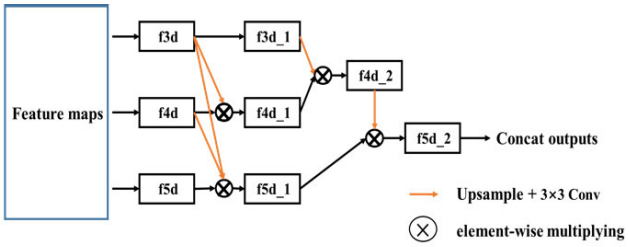

FIGURE 4. A schematic diagram of SPRFB module.

b: HOLISTIC ATTENTION MODULE (HAM)

The common feature extraction network is used to extract public and shallow feature information, and the multi-scale fusion decoder module is used to fuse feature information at multiple levels. However, how to better solve the problem of combining shallow feature information with feature maps that integrate multiple levels of feature information is a problem that needs to be solved urgently, so the holistic attention module came into being.

According to experience, the shallow feature information has rich detailed information, and the high-level abstract feature information has strong semantic information, but the edge information of the object is seriously lost. HAM uses low-level detailed information to recover the semantic loss of high-level edge information. Specifically, the operation flow of the HAM is: for multi-scale fusion feature information f_{id}, filtering by the Gaussian filtering and get the filtered result. Then the filtered result is normalized. Finally, each pixel takes the maximum value between the low-level feature map f_ni and the normalized result. The overall attention module is expressed by the following formula:

$$f_h = \max\left(f_{\min\_\max}\left(\mathrm{Conv}_g\left(f_{id}, k\right)\right), f_i\right) \tag{5}$$

Conv_g is a convolution operation with Gaussian function as a parameter, $f_{\min\_\max}$ is a regularization term, and the obtained prediction feature map is regularized to the range of [0,1], max() is the maximum value operating.

3) COARSE ALPHA BRANCH

This branch performs a coarse alpha regression task, which is different from the segmentation task in the previous stage. The difference lies in the following aspects:

- The trimap segmentation task is rough segmentation, and the alpha mask segmentation requires refined regression.
- The trimap segmentation task requires high-level semantic information to assist, and the alpha mask segmentation requires low-level features information to recover. Based on the above two differences, the network structure design at this stage needs to be adjusted accordingly.

With regard to the first difference, this stage requires richer detailed information to recover the lost detailed information, so this part introduces the first three layers of feature map information. In addition, another important part at this stage is the need for specific information fusion. For the trimap information output by the trimap segmentation network, it is necessary to extract the most concerned area in the task, that is, the unknown area. This area is the key part of the alpha regression. The main processing method is to perform specific information fusion on the output of the channel information of the unknown area, and the low-level feature information

$$\mathrm{attention}_{f_i} = f_i \odot (Up(UK)) \tag{6}$$

Up(·) represents an up-sampling operation with a ratio of 2^{i-1}. In this up-sampling operation, a bilinear interpolation method is used. UK represents the unknown area, f_i is the low-level feature information generated by the common feature extraction network, $\odot$ represents the element-wise multiplication operation, and finally generates specific fusion feature information $attention_{f_i}$.

$$\mathrm{CA} = \mathrm{MSD}\left(\left[\mathrm{attention}_{f_i}, Up(t)_i\right]\right), \quad i \in [1, \ldots, 3] \tag{7}$$

Up(·) represents an up-sampling operation with a ratio of 2^{i-1}. In this up-sampling operation, the nearest neighbor interpolation method is used. t represents the trimap, MSD represents the multi-scale fusion decoder module, and finally generates coarse alpha matte CA.

4) REFINE ALPHA BRANCH

Inspired by the two-stage design in DIM, this branch performs the refinement processing of the coarse alpha matte. In the coarse alpha regression stage, by visualizing the regression effect of coarse alpha, we find that the values in unknown area distribute in a range near 0.5. In order to solve this problem, this paper directly introduces a cascade branch. The branch design can be summarized by the following formula:

$$\mathrm{RA} = \mathrm{MSD}\left(\left[\mathrm{attention}_{f_i}, Up(t)_i, Up(CA)_i\right]\right), \quad i \in [1, \ldots, 3] \tag{8}$$

where Up(·) represents an up-sampling operation with a ratio of 2^{i-1}. In this up-sampling operation, the nearest neighbor interpolation method is used. t represents the trimap, CA represents the coarse alpha matte. MSD represents the multi-scale fusion decoder module.Finally the refine alpha matte RA is generated.

IV. RESULTS AND DISCUSSION

A. LIMITATIONS OF THE TRIMAP SEGMENTATION

In order to verify the deficiencies of the semantic segmentation in the trimap segmentation algorithm and elicit our solution, we first conducted a verification test. The following describes the details of the verification test with respect to the data, training process and results.

1) DATA PREPARATION

In this verification experiment, two types of label data are required, including high-precision trimap segmentation label images and coarse-precision trimap segmentation label images. In order to obtain a high-precision trimap segmentation image, this paper has made some improvements to the traditional trimap generation method. Generally speaking, the method of obtaining trimap is to rely on its alpha matte to perform random scale erosion and expansion operations. However, this approach will have many isolated cluster noises. In order to eliminate these noises, our method employs image connectivity and designs a noise area threshold to filter out these isolated cluster noises, so that a more high-precision trimap label can be obtained. To get the coarse trimap, 1/20 of the width and height of the image act as a random ratio of erode and expansion operations on high-precision trimap.

Also, the DeepLabV3+ model was also used to generate binary labels with only foreground and background. The foreground and background were distinguished by controlling the threshold to generate a saliency mask map. After that the saliency trimap and the saliency score map can be obtained simultaneously.

2) EXPERIMENT SETTING AND CONCLUSIONS

This experiment is used to verify whether a high-precision trimap output can be obtained when using a single segmentation network without additional prior information. The segmentation model adopts the most advanced DeepLabV3+, and five sets of verification experiments are designed. The only difference lies in the input, which are: 1) a single RGB image (RGB), 2) an RGB image with coarse trimap (RGB+CT), 3) an RGB image with a saliency mask map (RGB+SM), 4) an RGB image with a saliency trimap (RGB+ST), 5) an RGB image with a saliency score map (RGB+SS). The comparison of experimental results is shown in Table1.

Comparing the results, it is obvious that for a single image segmentation model with only RGB images as input, the final values of FgIoU, UKIoU, and BgIoU are all lower than the other four models, and the overall accuracy is also significantly lower than the other four models, which shows that it is difficult to obtain high-precision trimap output without adding additional prior information using a single segmentation network. However, after adding low-precision trimap prior information to a single segmentation network, the three values of FgIoU, UKIoU, and BgIoU all soared by 14%, 39%, and 15%, and the Acc soared to 98.28%. This phenomenon shows that only a single segmentation model is semantically ambiguous. In the task of high-precision segmentation of foreground and unknown region, additional prior information is needed, and the use of prior information can better segment the foreground and unknown region.During the experiment, it was found that generating saliency attention map (two classification) was much easier than trimap (three classification), and the saliency attention map could effectively guide the efficient generation of trimap.

Based on the hypothesis that the single-segment network can obtain high-precision trimap output with additional prior information, this paper further expands the experiment, to validate whether the additional information that needs to be added can be replaced with saliency mask information that only has foreground and background information. Therefore, the following experiments were carried out. The experimental results are shown in the last three rows of Table1.

Observing the experimental results, it is found that the additional prior information that adding the saliency mask reaches 96.35%, 92.07%, and 88.92%, which are all higher than the single RGB input model, which further proves that the single segmentation network does not add additional prior information. It is difficult to get high-precision trimap output. At the same time, the Acc of the experimental RGB+SS reached 96.35%, which is very close to the 98.28% experimental result of RGB+CT. It is proved that adding the prior information of the saliency mask can also achieve the segmentation task of generating high-precision trimap, and the acquisition of saliency mask information is simpler and more convenient than the acquisition of coarse trimap. Therefore, we propose a two-step strategy to first detect the saliency of foreground objects, and then obtain high-precision trimap segmentation images to solve the problem of high-precision trimap acquisition.

B. DATASETS

This paper synthesizes a personal image matting dataset. The training set contains 12471 foreground portrait images, and the test set contains 1239 foreground portrait images. These data are all from the network and are carefully labeled manually. Similar to the DIM [14] data set, background pictures are randomly selected from the COCO-2012 [30] data set for new data synthesis. The difference is that the work of this paper extends the DIM dataset, selecting background pictures from multiple scenes, and constructing a large background data set. Multiple scenes include indoor, car, street, and distant scenes, and multiple background switching make the synthesized data more authentic.

TABLE 1. Result of validation test.

model	input type	FgIoU/%	UkIoU/%	BgIoU/%	Acc/%
DeepLabv3+	RGB	84.59	14.03	82.04	2.79
DeepLabv3+	**RGB+CT**	**98.08**	**53.49**	**97.25**	**98.28**
DeepLabv3+	RGB+SM	97.60	28.06	97.01	96.35
DeepLabv3+	RGB+ST	94.93	21.24	94.47	92.07
DeepLabv3+	RGB+SS	95.59	23.52	94.94	88.92

TABLE 2. Comparison of trimap segmentation results.

model	input type	FgIoU/%	UkIoU/%	BgIoU/%	mIoU/%
DeepLabv3+	RGB	84.59	14.03	82.04	60.22
DeepLabv3+	RGB+CT	98.08	53.49	97.25	82.94
DeepLabv3+	RGB+SM	97.60	28.06	97.01	74.22
CSMNET+IS320	RGB	96.32	32.43	97.23	75.33
CSMNET+SSIM+IS320	RGB	98.03	51.21	96.24	81.83
CSMNET+SSIM+FA+AC+IS320	RGB	98.32	53.23	97.04	82.87
CSMNET+SSIM+FA+AC+IS512	RGB	97.46	56.25	95.45	83.05
CSMNET+SSIM+AC+IS320	RGB	98.28	55.98	**98.37**	84.21
CSMNET+SSIM+FA+IS320	RGB	98.45	54.56	98.03	83.68
CSMNET+SSIM+FA+IS512	**RGB**	**98.65**	**57.44**	97.68	**84.59**

TABLE 3. Comparison of alpha matte results.

model	input type	MSE	SAD
CSMNET+IS320	RGB	0.0743	0.2967
CSMNET+SSIM+IS320	RGB	0.0193	0.1880
CSMNET+SSIM+FA+AC+IS320	RGB	0.0150	0.1401
CSMNET+SSIM+FA+AC+IS512	RGB	0.0144	0.3065
CSMNET+SSIM+AC+IS320	RGB	0.0162	0.2056
CSMNET+SSIM+FA+IS320	RGB	0.0188	0.1516
CSMNET+SSIM+FA+IS512	**RGB**	**0.0118**	0.3252

C. PARAMETERS SETTING

In the model training, the PyTorch framework is used, with Adamax as the optimization method, and the weight decay is set to 5e-4. The learning rate lr is 0.0001, and the batch size is set to 16. In the process of initializing the model parameters, the common feature extraction network parameters are the pre-training parameters of VGG-16 in the ImageNet challenge. The parameters of other parts are set by the way of Kaiming initialization. For all training tasks, the iterative process sets epoch to 100. In terms of data augmentation, including random flipping, random angle rotation, the angle interval is $[-45°, 45°]$, and finally zooming to 320×320 and 512×512 fixed size for training.Staged training and end-to-end fine-tuning ensured the robustness of the model to the segmentation error.Because it is a cascading model of multiple subnetworks, we adopted a staged training manner during the training process. Specifically, we will train the first stage, and when it converges, fix its parameters to continue training the next stage. Finally, we will also conduct end-to-end overall training on the entire model.

D. LOSS FUNCTION

Since the overall network design is hierarchical, the training form of the entire model is end-to-end overall training. For each sub-task, a different loss function is used. The overall loss function is designed as the result of the proportional fusion of the sub-task loss functions.

$$\mathcal{L} = \lambda_0 \mathcal{L}_{\text{ssim}} + \lambda_1 \mathcal{L}_{\text{ce}} + \lambda_2 \mathcal{L}_{\text{a}} + \lambda_3 \mathcal{L}_{\text{c}} \tag{9}$$

This paper introduces structure similarity loss (SSIM) [32] to help the model better learn the structure information of the saliency object. On the whole, the loss considers the model loss at the block level, considering the local area adjacent to each pixel. At the same time, the loss can give more weight to the boundary of the saliency object, so it can assist the model to learn more refined object boundary contours.

$$l_{\text{ssim}} = 1 - \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)} \tag{10}$$

In the trimap segmentation stage, cross entropy loss is employed. On the other hand, the alpha regression task uses two loss functions including alpha predict loss and composition loss.

$$\mathcal{L}_\alpha^i = \sqrt{\left(\alpha_p^i - \alpha_g^i\right)^2 + \epsilon^2}, \alpha_p^i, \alpha_g^i \in [0, 1] \tag{11}$$

$$\mathcal{L}_c^i = \sqrt{\left(c_p^i - c_g^i\right)^2 + \epsilon^2} \tag{12}$$

E. RESULTS

This paper has conducted ablation experiments for the proposed modules. The following introduces the results of the experiment. The comparison of trimap segmentation and

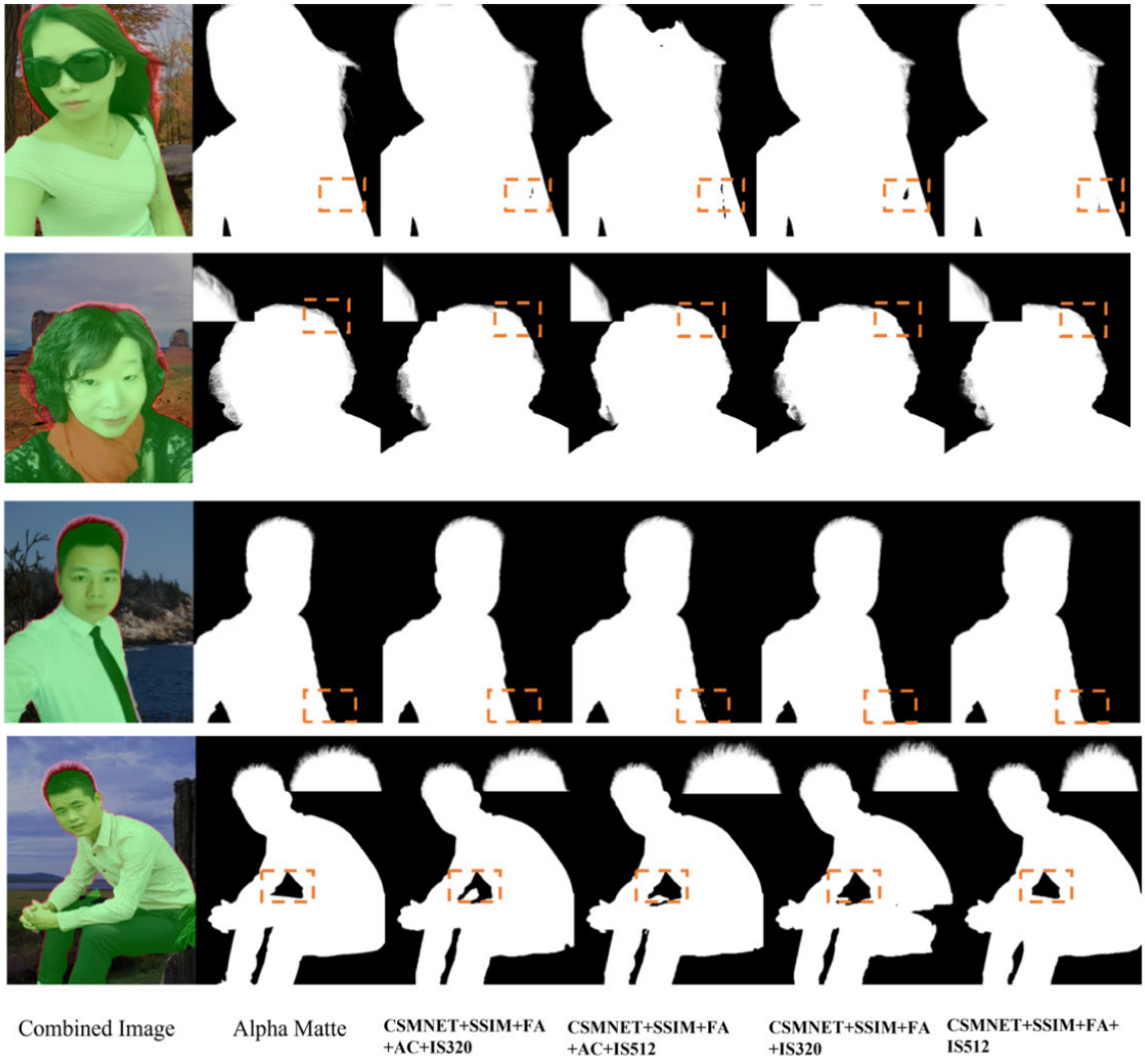

FIGURE 5. Results of CSMNet output. The picture in first column is a combined picture of the original training picture and the predicted trimap. The green part is the absolute foreground area, the pink part is the transition area, and the rest is the background area. The pictures in second column are listed as the corresponding alpha matte label images. Each picture in the last four columns represents the alpha matte output of different models. The models represented in the last four columns are CSMNET+SSIM+FA+AC+IS320, CSMNET+SSIM+FA+AC+IS512, CSMNET+SSIM+FA+IS320 and CSMNET+SSIM+FA+IS512, respectively.

alpha matte results can be seen in Table2. It is completely feasible and efficient to automatically generate high-precision trimap information by introducing saliency detection prior information with a single RGB input. In the trimap segmentation results of our methods, all experiments use the same RGB input and no additional prior information, the trimap three-category IoU and mean IoU are higher than those based on image segmentation. And the worst model is also increased by 15%. The IoU values of each category of most optimal model are 98.65%, 57.44%, 87.68%, and the mIoU value is 84.59%. Most of the evaluations are higher than single image segmentation model with low-precision trimap prior information model.

Joint loss function training helps improve performance. In the trimap segmentation effect, the IoU values of each category are 96.32%, 32.43%, 97.23% with model CSMNET+IS320, and the mean IoU value is 75.33%. The IoU values are 98.03%, 51.21%, 96.24% with model CSMNET+ SSIM + IS320, and the mIoU value is 81.83%, which is 2%, 19%, and 6% higher than the CSMNET+IS320 model. In the alpha matte regression effect, the MSE and SAD values with the CSMNET+IS320 model reached 0.0743 and 0.2967, respectively, and the MSE and SAD values with the CSMNET+SSIM+IS320 model reached 0.0193 and 0.1880, respectively, which were 0.6% and 11% lower than the CSMNET+IS320 model.

Holistic attention module helps to get a more accurate trimap, and then a more detailed alpha matte. The comparison results are listed in Table3, and the experiment is aimed at proving the validity of the FA module.

When the input resolution IS320 is fixed, comparing the model CSMNET+SSIM+FA+AC+IS320 and the model CSMNET+SSIM+AC+IS320, and the MSE and SAD values decrease by 0.11 and 0.06 respectively with FA model. Comparing the model CSMNET+SSIM+IS320 and CSMNET+SSIM+FA+IS320, the MSE and SAD values of the FA model decrease by 0.001 and 0.03, respectively. The experimental trends of the two editions are consistent, indicating that the Holistic attention module promotes the accurate segmentation of the alpha matte.

Larger input resolution makes it easier to get better alpha matte. As shown in Table3, the models correspond to IS320 and IS512 suffix names respectively. First, let's compare the results of trimap segmentation under different image resolution sizes, as shown in Table2. Comparing the CSMNET+SSIM+FA+AC+IS320 and CSMNET+SSIM+FA+AC+IS512 models, the IS512 model has a 3.02% improvement in the IoU value in the unknown area of the trimap, and the mean IoU is increased by 0.18%. Comparing CSMNET+SSIM+FA+IS320 and CSMNET+SSIM+FA+IS512 models, the IS512 model has a 2.88% improvement in the IoU value of the unknown area of the trimap, and an mean IoU improvement of 0.91%. Later we show the comparable results in alpha matte. The MSE and SAD values with the CSMNET+SSIM+FA+AC+IS320 model reached 0.0150 and 0.1401, respectively. The MSE and SAD values with the CSMNET+SSIM+FA+AC+IS512 model reached 0.0144 and 0.3065, respectively. The MSE and SAD values with the CSMNET+SSIM+FA+AC+IS320 model reached 0.0144 and 0.3065. The SAD values reached 0.0188 and 0.1516, and the MSE and SAD values with the CSMNET+SSIM+FA+IS512 model reached 0.0118 and 0.3252, respectively.

The model realizes end-to-end portrait matting and has excellent hair fineness. The visualization results can be seen in Fig.5. With respect to the pictures in column (a), this part selects the trimap output of model CSMNET+SSIM+FA+IS512. It can be found that the trimap has the characteristics of high accuracy and small unknown area. Secondly, observing the comparison chart of the alpha matte output of each model, the CSMNET+SSIM+FA+IS512 model's alpha matte is the most accurate. By comparison, it is found that the model is robust and can fit both difficult and simple samples normally. For example, portrait photos of boys and girls can achieve fine regression of the hair area. Moreover, it can be found that the generalization performance of the model is good. For large-pose data, such as the 3rd and 4th rows of pictures, the optimal model can also achieve accurate alpha matte

V. CONCLUSION

In this paper, we proposed cascaded segmented matting network (CSMNet), a new CNN architecture which learned trimap segmentation and alpha regression tasks in a collaborative way to achieve the end-to-end human image matting. Specifically, a shared encoder and two separate decoders are designed. Comprehensive empirical studies on above dataset demonstrate that CSMNet can produce a stable and accurate alpha matte without trimap and achieve an evaluation value that is comparable to the algorithm that requires trimap. Nowadays, short videos are developing well, and CSMNet has only been well processed for static data. Therefore, model lightweighting is the goal of future research, to make it more efficient to process images in videos.

REFERENCES

[1] T. Porter and T. Duff, "Compositing digital images," in *Proc. 11st Annu. Conf. Comput. Graph. Interact. Techn.*, 1984, pp. 253–259.
[2] A. Levin, D. Lischinski, and Y. Weiss, "A closed-form solution to natural image matting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 2, pp. 228–242, Feb. 2008.
[3] K. He, C. Rhemann, C. Rother, X. Tang, and J. Sun, "A global sampling method for alpha matting," in *Proc. CVPR*, Jun. 2011, pp. 2049–2056.
[4] Q. Chen, D. Li, and C. K. Tang, "KNN matting," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 9, pp. 2175–2188, Sep. 2013.
[5] Y. Aksoy, T. O. Aydin, and M. Pollefeys, "Designing effective inter-pixel information flow for natural image matting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 29–37.
[6] H. Chang, Q. Yang, and C. Pan, "An iterative Bayesian approach for digital matting," in *Proc. 18th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2006, pp. 122–125.
[7] Q. Chen, T. Ge, Y. Xu, Z. Zhang, X. Yang, and K. Gai, "Semantic human matting," in *Proc. 26th ACM Int. Conf. Multimedia*, Oct. 2018, pp. 618–626.
[8] V. Gupta and S. Raman, "Automatic trimap generation for image matting," in *Proc. Int. Conf. Signal Inf. Process. (IConSIP)*, Oct. 2016, pp. 1–5.
[9] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 640–651, Dec. 2015.
[10] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.
[11] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4353–4361.
[12] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
[13] J. Sun, J. Jia, C.-K. Tang, and H.-Y. Shum, "Poisson matting," *ACM Trans. Graph.*, vol. 23, no. 3, pp. 315–321, 2004.
[14] N. Xu, B. Price, S. Cohen, and T. Huang, "Deep image matting," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2970–2979.
[15] J. Tang, Y. Aksoy, C. Oztireli, M. Gross, and T. O. Aydin, "Learning-based sampling for natural image matting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3055–3063.
[16] Q. Hou and F. Liu, "Context-aware image matting for simultaneous foreground and alpha estimation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 4130–4139.
[17] S. Cai, X. Zhang, H. Fan, H. Huang, J. Liu, J. Liu, J. Liu, J. Wang, and J. Sun, "Disentangled image matting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8819–8828.
[18] H. Lu, Y. Dai, C. Shen, and S. Xu, "Indices matter: Learning to index for deep image matting," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3266–3275.
[19] S. Sengupta, V. Jayaram, B. Curless, S. M. Seitz, and I. Kemelmacher-Shlizerman, "Background matting: The world is your green screen," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2291–2300.
[20] Z. Ke, K. Li, Y. Zhou, Q. Wu, X. Mao, Q. Yan, and R. W. H. Lau, "Is a green screen really necessary for real-time portrait matting?" 2020, *arXiv:2011.11961*.
[21] X. Yang, K. Xu, S. Chen, S. He, and B. Y. Yin, "Active matting," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 4590–4600.

[22] Y. Zhang, L. Gong, L. Fan, P. Ren, Q. Huang, H. Bao, and W. Xu, "A late fusion CNN for digital matting," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7469–7478.
[23] O. Wang, J. Finger, Q. Yang, J. Davis, and R. Yang, "Automatic natural video matting with depth," in *Proc. 15th Pacific Conf. Comput. Graph. Appl. (PG)*, Oct. 2007, pp. 469–472.
[24] N. Joshi, W. Matusik, and S. Avidan, "Natural video matting using camera arrays," *ACM Trans. Graph.*, vol. 25, no. 3, pp. 779–786, Jul. 2006.
[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[26] J. Huang, Z. Zhu, and G. Huang, "Multi-stage HRNet: Multiple stage high-resolution network for human pose estimation," 2019, *arXiv:1910.05901*.
[27] S. Liu and D. Huang, "Receptive field block net for accurate and fast object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 385–400.
[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[29] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-V4, inception-resnet and the impact of residual connections on learning," in *Proc. AAAI Conf. Artif. Intell.*, 2017, vol. 31, no. 1, pp. 4278–4284.
[30] T. Lin, M. Maire, and S. Belongie, "Microsoft COCO: Common objects in context," in *Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2014, pp. 740–755.
[31] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7479–7489.

BO LIU (Senior Member, IEEE) received the B.S. degree from the Department of Automation, Beijing Institute of Technology, Beijing, China, and the M.S. and Ph.D. degrees from the Department of Automation, Tsinghua University, Beijing, in 2003 and 2008, respectively. She worked with the NEC Laboratory China as a Researcher, from 2008 to 2010 and from 2013 to 2015. She was a Research Professional with the Computation Institute, The University of Chicago, Chicago, IL, USA, and Mathematics and Computer Science Division, Argonne National Laboratory, Lemont, IL, USA, from 2011 to 2012. She joined the Beijing University of Technology, in 2015, as an Associate Professor. Her current research interests include big data, data mining, machine learning, cloud computing, scientific workflow, semantic web, and ontology reasoning.

HAIPENG JING is currently pursuing the Graduate degree with the School of Software Engineering, Beijing University of Technology, Beijing, China. His current research interests include big data, semantic segmentation, and machine learning.

GUANGZHI QU (Senior Member, IEEE) received the B.E. and M.E. degrees from the Department of Computer Science and Engineering, Beihang University, and the Ph.D. degree in computer engineering from The University of Arizona, in 2005. He joined the Computer Science and Engineering Department, Oakland University, in 2007, where he is currently an Associate Professor. His research interests include data mining, machine learning, operating systems, and program analysis. He was the Conference Co-Chair of the 2014 International Conference on Machine Learning and Applications (ICMLA).

HANS W. GUESGEN received the doctoral degree in computer science from the University of Kaiserslautern, and the Habilitation degree in computer science from the University of Hamburg, Germany. He is currently a Professor in computer science with Massey University, New Zealand. His research interests include smart environments, ambient intelligence (ubiquitous computing with artificial intelligence), knowledge representation, and spatio-temporal reasoning, with more than 100 refereed papers in these areas. He is also a Senior Member of the Association for the Advancement of Artificial Intelligence (AAAI) and a Honorary Fellow of the Munich University of Applied Sciences. He has been a member of the programme committees of more than 70 international conferences and workshops, and has served as a Referee for the Australian Research Council, the U.S. National Science Foundation, the NZ Foundation for Research Science & Technology, and more than 70 international journals and conferences.