IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 29 November 2022, accepted 28 December 2022, date of publication 30 December 2022,
date of current version 5 January 2023.

Digital Object Identifier 10.1109/ACCESS.2022.3233419

RESEARCH ARTICLE

CSU-Net: Contour Semantic Segmentation Self-Enhancement for Human Head Detection

MOHAMED CHOUAI[1,2] AND PETR DOLEZEL[2], (Member, IEEE)
[1]Alfred Wegener Institute, 27515 Bremerhaven, Germany
[2]Faculty of Electrical Engineering and Informatics, University of Pardubice, 53210 Pardubice, Czech Republic

Corresponding author: Petr Dolezel (petr.dolezel@upce.cz)

This work was supported by the Programme INTER-EXCELLENCE (LTAIN19100) Funds of the Ministry of Education, Youth and Sports, Czech Republic, Artificial Intelligence Enabled Smart Contactless Technology Development for Smart Fencing, under Project LTAIN19100.

ABSTRACT The computer vision community has made tremendous progress in solving a variety of semantic image understanding tasks, such as classification and segmentation. With the advancement of imaging technology and hardware, image semantic segmentation, through the use of deep learning, is among the most common topics which have been worked on in the last decade. However, image semantic segmentation suffers from several drawbacks such as insufficient detection of object boundaries. In this study, we present a new convolutional neural network architecture called CSU-Net that aims to self-enhance the results of semantic segmentation. The proposed model consists of two strongly concatenated encoder-decoder blocks. With this design, we reduced requirements on computing power and memory size to decrease costs and increase the training/prediction speed. This study also demonstrates the advantage of the proposed system for small training data sets. The proposed approach has been implemented on our private dataset, as well as on a publicly available dataset. A comparative analysis was carried out with four popular segmentation models and three other recently introduced architectures to show the efficiency of the proposed system. CSU-Net outperformed the other competing neural networks that we considered for the comparative study. As an example, it succeeded in improving the traditional U-Net result by approximately 50% in mean Intersection over Union (mIoU) for both tested datasets. Based on our experience, the CSU-Net can improve results of semantic segmentation in many applications.

INDEX TERMS Safety systems, head detection, head counting, semantic segmentation, self-enhancement.

I. INTRODUCTION

Semantic segmentation has attracted a lot of attention from the computer vision community in numerous applications for many years. Semantic segmentation is basically a building block that allows understanding of the scene. By densely classifying all the pixels of a scene image, it is possible to construct abstract representations of objects and their shapes.

Deep convolutional neural networks (DCNNs) have been driving significant advances in semantic image segmentation due to their powerful feature representation for image processing. However, their performance in preserving object boundaries is still not satisfactory. This phenomenon is described very explicitly in a recent study [1]. Here, the authors point out that although semantic image segmentation methods using DCNNs provide impressive results, most of them neglect the long-distance dependencies between inner objects and boundaries. Another source pointing out this problem is survey [2]. The authors here argue that the outputs from the final layer of DCNNs are not sufficiently localized for accurate object boundaries due to their invariance properties. This makes precise boundary recovery of semantic segmentation an academic challenge.

Person recognition and detection is one of the applications that can be solved using semantic segmentation. One of the highly demanded practical applications is the detection of people in scenes captured by an RGB camera. With this capability, the flow of people or crowd density in the monitored area can be identified and analyzed, bringing many benefits to everyday life. A typical example is counting

The associate editor coordinating the review of this manuscript and approving it for publication was Diego Oliva.

the exact number of passengers entering and exiting public transport [3], which enables a forecast of passenger flows [4], the planning of transport schedules [5], and the monitoring of the loading of transport vehicles [6]. A second benefit is tracking people in video, especially in surveillance systems, to determine a person path and analyze their movement [7]. A third benefit is in stores and malls where this type of technology is used to collect data and improve business strategy [8].

Person detection is a very challenging environment for autonomous systems because of the dynamics of observed objects (pedestrian detection, behavior identification and prediction, etc.). Besides, most object locations in these scenes are very unpredictable; for example, it is possible to find a person in the middle of the image as well as in the corner of the image. Additionally, people's body height and thickness differ from several pixels to most of the image. Moreover, especially in crowded images, people in the scene are partially visible or only the upper part of their bodies can be observed. As a consequence, the assumption to take advantage of location information in the form of a digital high-definition map in perception modules seems unreasonable. The erroneous detection of a person can occur due to the following reasons: the lack of richness of the dataset, the low quality of the images, the overlap of several persons, different sizes of the objects, and processing every single image in a video sequence separately without any relation to the adjacent images.

In this study, we propose a new neural network semantic segmentation architecture to detect multiple persons from an RGB image - the key feature of this detection method is the focus on people's heads. This work is based on a long-term collaboration between the authors' team and industry partners. The achieved results follow the previous publication [9], where the problem of classification and detection of persons in visual data was solved using HOG descriptors; and the publication [10], where an improved DeepLabv3+ semantic segmentation approach was proposed to detect a human head in an RGB image.

Consequently, a novel convolutional neural network architecture, based on two strongly concatenated shallow U-Net networks, is proposed. The first block aims to apply semantic segmentation, the second one is expected to self-enhance the result gained from the first network. By forcing the enhancement block to improve the output of the main block, the system has the ability to allow additional convolutional mappings to reflect more information about the original input image when generating features of a higher level. The motivation for using the enhancement block is the ability to more accurately detect the boundaries of the objects of interest, i.e, heads, especially in crowded and poorly illuminated areas. We believe that this improvement will bring better accuracy in person detection, in particular, preventing multiple persons from merging into one and mistaking the person for another object.

The proposed method was tested in several experiments and it is compared with other similar methods to verify its efficiency. The purposes of the experiments are to prove the efficiency of our method by answering the following questions: (1) can the model not only improve the representation power of features but also enhance the object contour performance? (2) is the model competitive in terms of computational cost and memory required to infer? and (3) why does the concatenation between the two networks improve the results?

The approach has been implemented in our private dataset as well as in the public dataset PAMELA UANDES [11]. With a 47% improvement in mIoU to the original U-Net architecture, the result demonstrated that the proposed system could be an efficient way to build a deep neural network model for semantic segmentation. We believe this can be applied not only for human head detection but also in other semantic segmentation applications.

The rest of this article is organized as follows. In Section II, we provide an overview of the research that has been carried out in previous years in the area of improving semantic segmentation. Section III gives a description of the CSU-Net network architecture and the training procedure. The benchmarking study is presented in Section IV. Section V shows the experimental results in which we describe the used datasets, the evaluation metrics, and the implementation details. We also analyze, interpret, and discuss our results. Finally, Section VI closes the article with conclusions and future work.

II. IMPROVING SEMANTIC SEGMENTATION

Semantic segmentation problems have been investigated using many different neural network architectures. Convolutional Neural Networks (CNNs) are, these days, at the center of attention when considering applications in semantic segmentation [12]. Most CNN architectures are based on an encoder-decoder design, such as [13], [14]. Moreover, skip connections [15] and dilated convolutions [16] preserve details in the segmentation, and spatial pyramid pooling [17] or global pyramid pooling [18] aggregate different scales to exploit spatial context information. Taken together, the CNNs thus provide very accurate results for a variety of complex semantic segmentation tasks.

Nevertheless, image semantic segmentation remains an unsolved challenging problem in the field of low-level computer vision, especially for complex scene understanding due to limited receptive fields and short-range information [1]. Several attempts have been made to address the problem.

Hoyer et al. [19] introduced three strategies capable of leveraging the knowledge learned from self-supervised depth estimation to improve semantic segmentation in both the semi-supervised and the fully supervised setting. Li et al. [20], developed a framework to improve the semantic segmentation results by decoupling features into the body and the edge parts to handle inner object consistency and fine-grained boundaries jointly. Equally, Yin et al. [21] proposed a model that links a branch of edge features and a branch of semantic features to ensure consistency between these feature values. This model clearly improves the precision of

segmentation results. Moreover, Fabio Yuluaga et al. [22] presented the cooperation between two inception blocks with an inner skip connection inside the blocks.

On the other hand, Zhu et al. [23] proposed joint image-label propagation and boundary label relaxation to improve the semantic segmentation results. They scaled up training sets and mitigated label noise during training to achieve that goal. Further, Pasad et al. [24] used models predicting depth, egomotion, and camera intrinsics, to provide additional supervision to a semantic segmentation model through spatio-temporal consistency constraints and improve the result. In addition, Zanjani et al. [25] focused on the scene dynamic information between the streams of frames in a video in order to increase the accuracy of semantic image segmentation. For this purpose, they proposed a method for integrating short-term temporal information with structural scene information by using a conditional random field.

Furthermore, publications inspired by the original U-Net architecture continue to appear. Colman et al. [26] introduced a deep residual bottleneck to the U-Net, and Ange Lou [27] added the dilated channel-wise CNN module and simplified the U-shaped layout in order to get a lightweight but efficient model.

Loukkal et al. [28] presented three different approaches to inject location information in semantic segmentation CNNs applied to urban scenes. Divecha et al. [29] proposed an improvement system of semantic segmentation for autonomous vehicles using synthetic images. To do so, they took advantage of an unlimited source of annotated data in virtual environments, and then transformed the data to have a more photo-realistic look, which matched their real-world counterparts. Furthermore, Huang et al. [30] developed an encoder-decoder network to solve the issue of the mutually exclusive relationship between the semantic response value and the semantics of object/component.

Additionally, Huang et al. [31] investigated the utility of motion boundaries in the improvement of semantic segmentation. This approach avoids extracting feature maps based on a high-to-low encoder, which may easily lose important shape and boundary details. Wu et al. [32] proposed a system that can maintain high-resolution features using a relatively shallow and parallel network structure. Also, Zhang et al. [33] proposed a probabilistic superpixel-based dense conditional random field model to refine label assignments as a post-processing optimization method, in order to preserve object boundaries in the semantic segmentation.

Niu [34] proposed a semantic segmentation method for remote sensing images based on CNN and mask generation. They used the boundary box as an initial foreground segmentation profile, and a multilayer feature of the convolutional neural network to provide the edge information of the foreground object. He et al. [35] designed advanced network architectures to incorporate a more suitable context and extract more representative features by developing an adversarial feature generator. Dong [36] proposed a method of image semantic segmentation using a generative adversarial network (GAN) combined with the ERFNet model in order to address the problems of insufficient segmentation of small-scale targets and weak anti-noise ability.

Zhu et al. [37] used a self-training paradigm with a semi-supervised approach to improve the semantic segmentation. Specifically, they trained a model on labeled data, and then generated pseudo labels on a large set of unlabeled data. To improve the semantic segmentation result, Shen et al. [38] proposed a novel region attention network for modeling the dependency between the object regions in order to compute the contextual representations. Moreover, Farsi and Mohammadzadeh [39] developed a model to reduce the amount of computational cost and memory required, and to increase speed/accuracy by proposing a 15-times reduction in the number of parameters of a SegNet network.

Wu et al. [40] proposed extra learning of dilated affinity information in the DeepLab v3+ training to help the learning process and to refine it with a fast affinity propagation post-processing, which exploits the extra information generated by the network. As well, Tran et al. [41], developed a system using focal loss, poly learning rate, and context module to improve the robustness of semantic segmentation for satellite images. Finally, Gritzner and Ostermann [42] mitigated the problem of an insufficient amount of training examples by using labeled source domain training examples and unlabeled target domain images to train a model.

The survey of the state-of-the-art methods shows that in this field there are different deep learning approaches applied to improve the outcome of semantic segmentation. However, there is no application developed for a self-improvement semantic segmentation system, which is our focus for this study.

III. METHODOLOGY

A. NETWORK ARCHITECTURE

The network design is illustrated in Figure 1. Our model consists of two blocks: the main block and the enhancement block. Both blocks use an encoder-decoder architecture inspired by the U-Net model. The main block takes an RGB image as input and provides a single depth mask image, where head contours are depicted in white. The input to the enhancement block is the input RGB image, concatenated with the mask obtained from the main block, to generate the enhanced mask in its output. The network architecture is detailed in Figure 2.

Our model is a convolutional neural network architecture based on two shallow U-shaped strongly concatenated networks. The first one aims to apply semantic segmentation, the second one to self-improve the result from the first one. By forcing the enhancement block to improve the output of the main block, the system will have the ability to allow additional convolutional mappings to reflect more information about the original input image when generating features of a higher level. This constitutes the advantage of this system.

The core idea of this paper is to utilize features typically extracted by fully convolutional network architecture during semantic segmentation, and then to increase the predicted

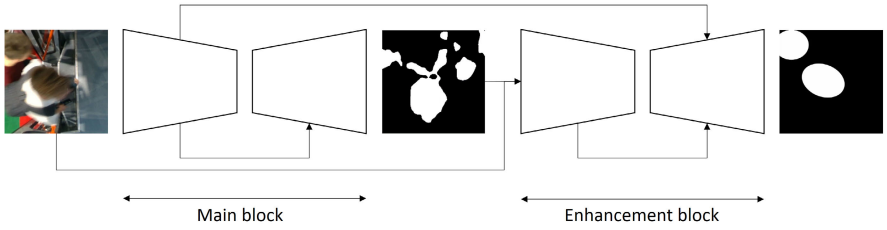


FIGURE 1. CSU-Net design. It is composed of two blocks strongly concatenated to each other.

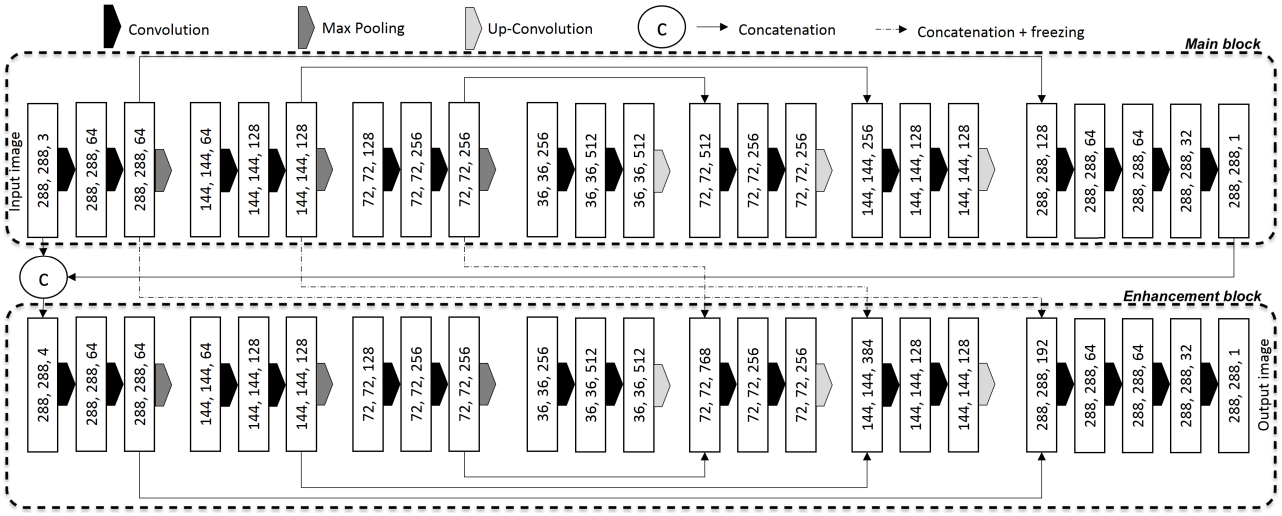


FIGURE 2. Block diagram describing the architecture of the CSU-Net.

segmentation result to match with a ground truth boundary. Specifically, we propose a series of concatenation combinations between a layer of the encoder of the main block and the corresponding layer of the decoder in the enhancement block. The concatenation is applied in order to provide more detailed feature information about the input data to the enhancement block, which is responsible for self-enhancing the result; see the dashed lines in Figure 2. This concatenation is applied with freezing of the convolutional layer weights and biases of the main block encoder to ensure the constancy of the features acquired during the training of the main block.

B. TRAINING PROCEDURE

We denote by x_i the input image and by q_i the ground truth, where $x_i \in R^{288\times288\times3}$ and $q_i \in R^{288\times288}$. We assume that both x_i and q_i are normalized in the interval [0,1]. The ground truth preparation will be explained in Section V-A.

The main and enhancement blocks of our architecture are defined with the functions $T_m(x_i, \theta_m)$ and $T_e(x_i, \theta_e)$ respectively. All trainable parameters of the main block are defined with vector θ_m, and similarly with θ_e for the enhancement block.

We train our network using a set of corresponding images and ground truth $x_i, q_i, i = 1, \ldots, N$, with the loss function based on the binary cross-entropy loss, since the problem is analogous to semantic segmentation, i.e., pixel-wise classification [43]. The loss is presented in eq. (1):

$$Loss(\theta) = -\frac{1}{N}\sum_{i=1}^{N} y_{m,i}\log(p_{m,i}) + (1-y_{m,i})\log(1-p_{m,i})) - \alpha\frac{1}{N}\sum_{i=1}^{N} y_{e,i}\log(p_{e,i}) + (1-y_{e,i})\log(1-p_{e,i})) \quad (1)$$

where $y_{m,i}$ is a binary indicator (0 or 1) if $T_m(x_i, \theta_m)$ provides the correct classification for input image pixel x_i. The same applies for $y_{e,i}$ and $T_e(x_i, \theta_e)$. Furthermore, $p_{m,i}$ is the predicted probability of the classification using the main block, and $p_{e,i}$ is the predicted probability of the classification using the enhancement block. Lastly, α is the hyperparameter that weights the importance of the main and enhancement terms in the loss function. The optimal value of α was empirically found on validation data, and was set to $\alpha = 1$ by evaluating its effect on the network performance metrics.

IV. COMPARATIVE ANALYSIS

Deep learning convolutional neural networks are widely used to solve semantic segmentation problems. It allows more complex tasks to be tackled through image segmentation.

To show the efficiency of the proposed system, we have chosen to compare it with four very popular segmentation models and three promising recently published architectures.

The original articles, source codes and most important features are summarized in Table 1.

A. U-NET

With a "U" shape, the U-Net architecture is symmetrical and its operation is somewhat similar to autoencoders. It can be reduced to three main parts: the contraction path (downsampling), the bottleneck, and the expansion (upsampling) path. The encoder portion of the neural network compresses the input into a latent space representation, then a decoder constructs the output from the compressed or encoded representation. With the U-Net, skip connections are used to transfer fine-grained information from the low-level layers of the analysis path to the high-level layers of the synthesis path. This information is needed to generate reconstructions that have fine detail [15].

B. PSPNet

The architecture of the Pyramid Scene Parsing Network (PSPNet) is based on the global pyramid pooling feature, which provides additional contextual information. First, the feature maps are obtained by a base network (ResNet101, DeepLab, etc.). A pyramid analysis module is applied to collect different representations of sub-regions. Convolution is applied to the maps of grouped features. Next, all feature maps are oversampled to a common scale and concatenated to form the final feature representation. Finally, the representation is introduced into a convolutional layer to obtain the final prediction per pixel [18].

C. LinkNet

Similarly to other segmentation architectures, LinkNet uses the encoding-decoding strategy. The issue on this approach is upsampling this feature map to the original resolution and preserving the categorization of the pixels. In LinkNet, the input of each encoder layer is also provided to the output of the corresponding decoder. Using these connections, the lost spatial information is recovered and can be therefore used by the decoder and its upsampling operations [44].

D. FPN

The idea of Feature Pyramid Network (FPN) is to adopt a strategy of hierarchical prediction to achieve the goal of complementary advantages. FPN consists of two paths. The first path is the usual convolutional network for feature extraction. As the signal is propagated through the network, the spatial resolution decreases. With more high-level structures detected, the semantic value of each layer increases [17].

E. IRUNet

The model architecture is designed using the cooperation between two inception blocks. In order to reduce the vanishing gradient problem, the network architecture is built wide rather than deep, with a strategic positioning of skip connections. Two different types of skip connections are applied to provide an alternative gradient path in backpropagation.

TABLE 1. Description of the models selected for the comparative analysis.

Model	Model adapted from	Keras imple-mentation	Important characteristics
U-Net	[15]	[45]	U-shaped, skip connections between encoder and decoder
LinkNet	[44]	[46]	Input of each encoder layer bypassed to output of its corresponding decoder, fewer parameters of decoder part
PSPNet	[18]	[46]	Global pyramid pooling feature - provides additional contextual information
FPN	[17]	[46]	Top-down architecture with lateral connections to build semantic feature maps at all scales (multi-scale, pyramidal hierarchy)
IRUNet	[22]	[47]	Cooperation between two inception blocks, two types of skip connections
DR-Unet104	[26]	[48]	Deep residual bottleneck block, residual blocks in decoder, dropout after each convolution layer
CFPNet-M	[27]	[49]	Feature pyramid channel, simplified the U-shaped layout, lightweight

The first type of skip connection is positioned between the encoder and the decoder part similarly to U-Net. The second type is situated inside the two inception blocks. A detailed explanation of the architecture can be found in [22].

F. DR-Unet104

Deep residual U-Net with 104 convolutional layers (DR-Unet104) is based on the U-Net, but it brings multiple additions. It uses deep residual blocks in the decoder part of the architecture, implies a specific version of the bottleneck residual block, and adds dropout after each convolution block stack [26].

G. CFPNet-M

Channel-wise Feature Pyramid Network for Medicine (CFPNet-M) is a very lightweight architecture proposed for various biomedical applications. It implements a feature pyramid channel to a U-shaped architecture. It is expected to show competitive performances with great advantages of much fewer parameters and smaller model file size [27].

H. ABLATION STUDY

Moreover, a series of ablation experiments was performed to explore the efficiency and robustness of the proposed CSU-Net.

First, the performance of the main block itself from the CSU-Net architecture was explored. Explicitly, we considered the output of the main block as if it were the output of the overall model. This model is referred to as SU-Net.

Next, the effect the concatenation between the main block and the enhancement block (the dashed lines in Figure 2) was examined. Specifically, a series of concatenation combinations between a layer of the encoder of the main block and the corresponding layer of the decoder in the enhancement block were disconnected, and both blocks were trained independently. This model is referred to as 2SU-Net.

V. EXPERIMENTAL WORK

A. DATASETS

The proposed CSU-net was tested on the private and PAMELA UANDES dataset [11].

1) PRIVATE DATASET

This dataset is prepared from scratch (from acquisition over preprocessing to manual labeling of data). Here, we captured videos from different locations with the RGB-Depth RealSense D435 camera [50]. Out of these videos, we obtained a sequence of images, from which the images were randomly selected until a sufficient quantity was obtained. RGB images were only used for the purposes of this dataset, which consists of 7,000 images taken from seven locations. Figure 3 shows some examples taken from this dataset.

2) PAMELA UANDES DATASET

PAMELA UANDES dataset [11], which we used in this study, is composed of a monocular camera recording videos looking down on passengers alighting/boarding a metropolitan train, as it is shown in Figure 4. Ground truth data of this dataset were prepared manually.

We preprocessed this dataset before working on it (we obtained 11,315 images at the end of this step). We considered the following points when preparing the images:

- Cropping the original video to take only the region of interest and to decrease the calculation time during the training of the models.
- The need to take consecutive images in pieces so that the tracking can be applied in future work.

Eventually, two specific datasets (private and PAMELA UANDES dataset) were annotated in a specific way. In order to simplify the labeling process, it was reasonable to approximate the positions and shapes of heads using an appropriate geometric formation. Considering the elliptic shape of heads, ellipses of suitable size were implemented. An example of the labeling process is illustrated in Figure 5. We view the elliptical shape of the heads in ground truth images as an acceptable compromise between annotation accuracy and the difficulty of manual dataset creation.

Note that there is a large difference in the complexity of the two selected datasets. While PAMELA UANDES dataset contains only scenes from a single environment, and the persons are always at approximately the same distance from the camera sensor, the private dataset contains a range of environments and distances, and is taken under changing lighting conditions.

FIGURE 3. Examples taken from the first dataset.

B. EVALUATION METRICS

The evaluation was made according to the following criteria. All the scores are expressed as a percentage (%):

a: MEAN ACCURACY

Mean accuracy allows to indicate the percentage of correctly identified pixels for all classes. I can be calculated by eq. (2).

$$MC = \frac{1}{M}\sum_{c=1}^{M}\frac{TP_c}{TP_c + FN_c}, \tag{2}$$

where TP and FN represent true positive and false-negative errors respectively. M is the number of classes.

FIGURE 4. Example taken from the second dataset in which figure (a) is before and (b) is after the preprocessing of the dataset.

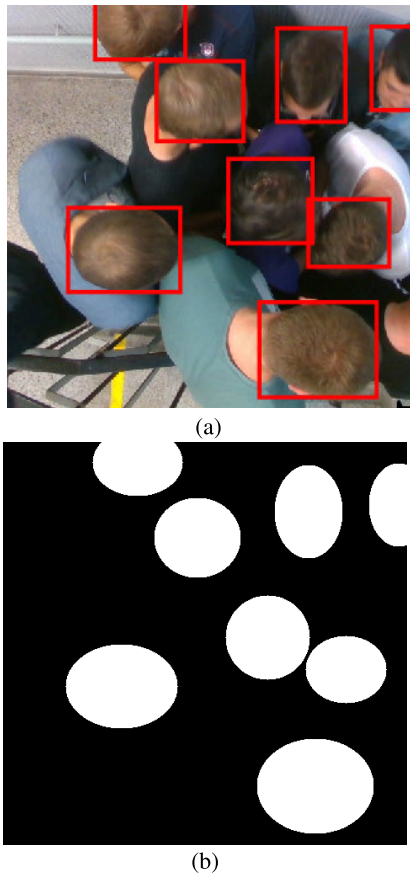


FIGURE 5. Preparation of the ground truth. (a) and (b) represent an example of: original bounding boxes superimposed on its raw image and the extracted ground truth respectively.

b: MEAN IoU

The Mean IoU is the average IoU score of all classes in all images. IoU is used to establish a measure of statistical precision that penalizes false positives. IoU is the ratio of correctly classified pixels to the total number of ground truths and predicted pixels in a class. The mean IoU is calculated by eq. (3).

$$MI = \frac{1}{M}\sum_{c=1}^{M}\frac{TP_c}{TP_c + FP_c + FN_c}, \tag{3}$$

where FP refers the false positive error.

c: MEAN BOUNDARY F1 SCORE

Mean boundary F1 (called as well BF Score) is the mean BF score of all classes in all images. Mean BF Score is the average BF score of the overall images of that class. BF score is a value in the range [0, 1]. A score of 1 means that the contours of the corresponding class objects, in prediction and in ground truth, match perfectly. The Mean boundary F1 score can be determined by eq. (4).

$$MBF = \frac{1}{M}\sum_{c=1}^{M}\frac{TP_c}{TP_c + \frac{1}{2}(FP_c + FN_c)}. \tag{4}$$

TABLE 2. Relevant features of the tested models.

Model	Trainable params	Memory size	Relative response time
U-Net	31,378,945	122 MB	95.69%
CSU-Net	15,566,402	60 MB	100%
LinkNet	20,325,137	79 MB	73.24%
PSPNet	10,101,589	39 MB	72.28%
FPN	17,595,605	68 MB	89.08%
IRUNet	7,810,689	30 MB	129.61%
DR-Unet104	76,516,529	300 MB	122.13%
CFPNet-M	654,279	4 MB	99.84%

d: RELATIVE RESPONSE TIME

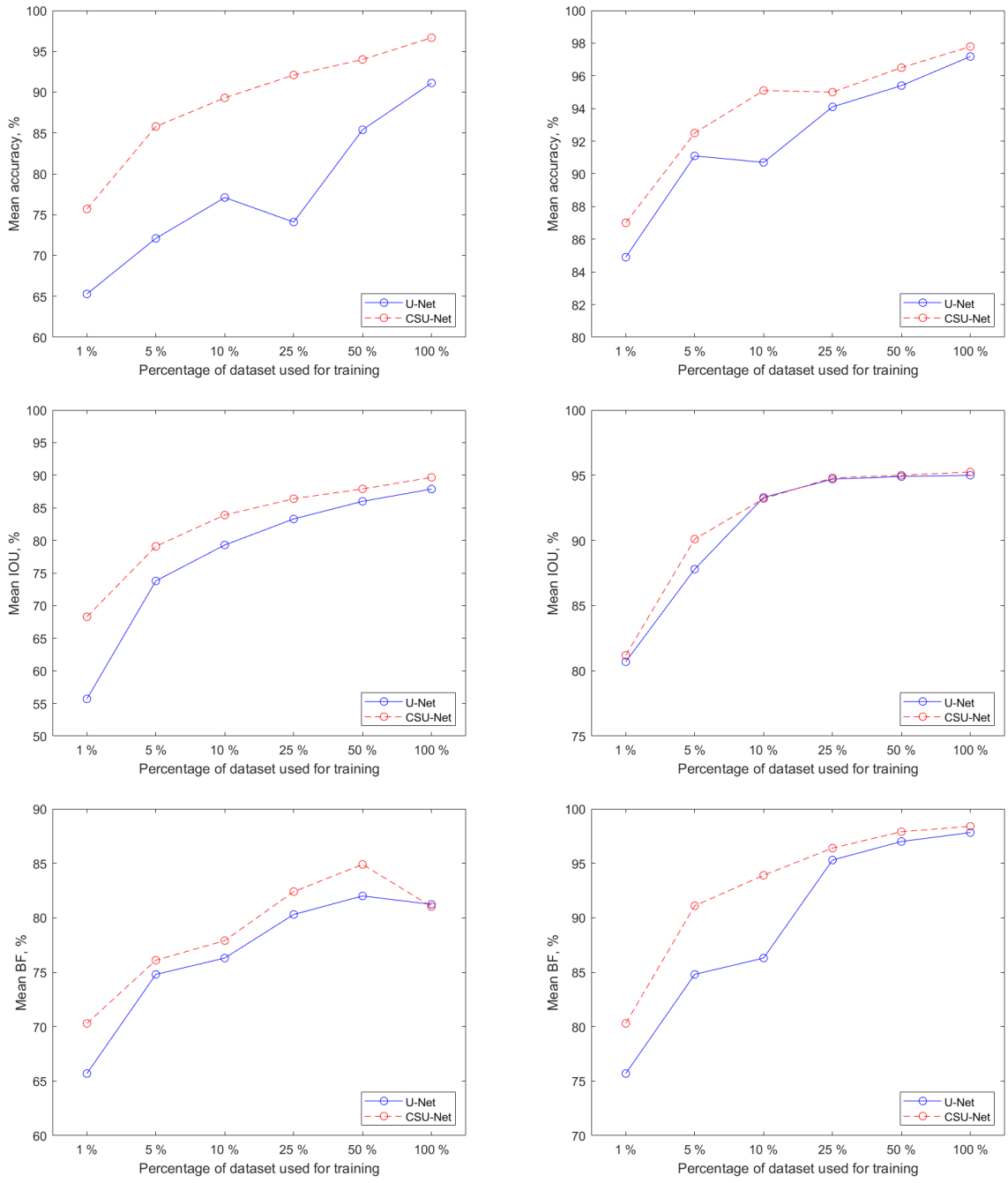
The relative response time of each considered model is defined as follows.

$$\tau = \frac{t_A}{t_{CSU}}, \tag{5}$$

where t_A is the response time of the considered model, and t_{CSU} is the response time of the proposed CSU-Net using the implementation conditions described in Section V-C.

C. IMPLEMENTATION DETAILS

The models were trained from scratch using the Adam optimizer. 180 epochs were used, along with a mini-batch size of 8, a learning rate of 0.001, and an L2 regularization



(a) Private dataset (b) UANDES dataset

FIGURE 6. **Performance of CSU-Net compared with U-Net in terms of mean accuracy, mean IoU and mean BF score with respect to the fraction of the amount of data used to train the models.**

factor of 0.0005. All training algorithms were performed using Keras/Python on a GPU of NVIDIA Quadro P5000 graphic card running on an operating system Windows.

Data augmentation was used for our application to improve network accuracy by randomly transforming the original data during training. For our application, random horizontal/vertical reflection, left/right random reflection, random X / Y translation of +/- 10 pixels, and random rotation were used for data augmentation.

D. RESULTS AND DISCUSSION

The objective of our study was to develop a system capable of enhancing the performance of state-of-the-art models. We have applied our proposed system and evaluated it by several metrics: Accuracy, IoU, and BF score. We used the boundary F1 (BF) score because it gives us a metric that tends to correlate better with human qualitative rating than the IoU metric. Moreover, the BF score is more significant than accuracy because the number of true negatives is not

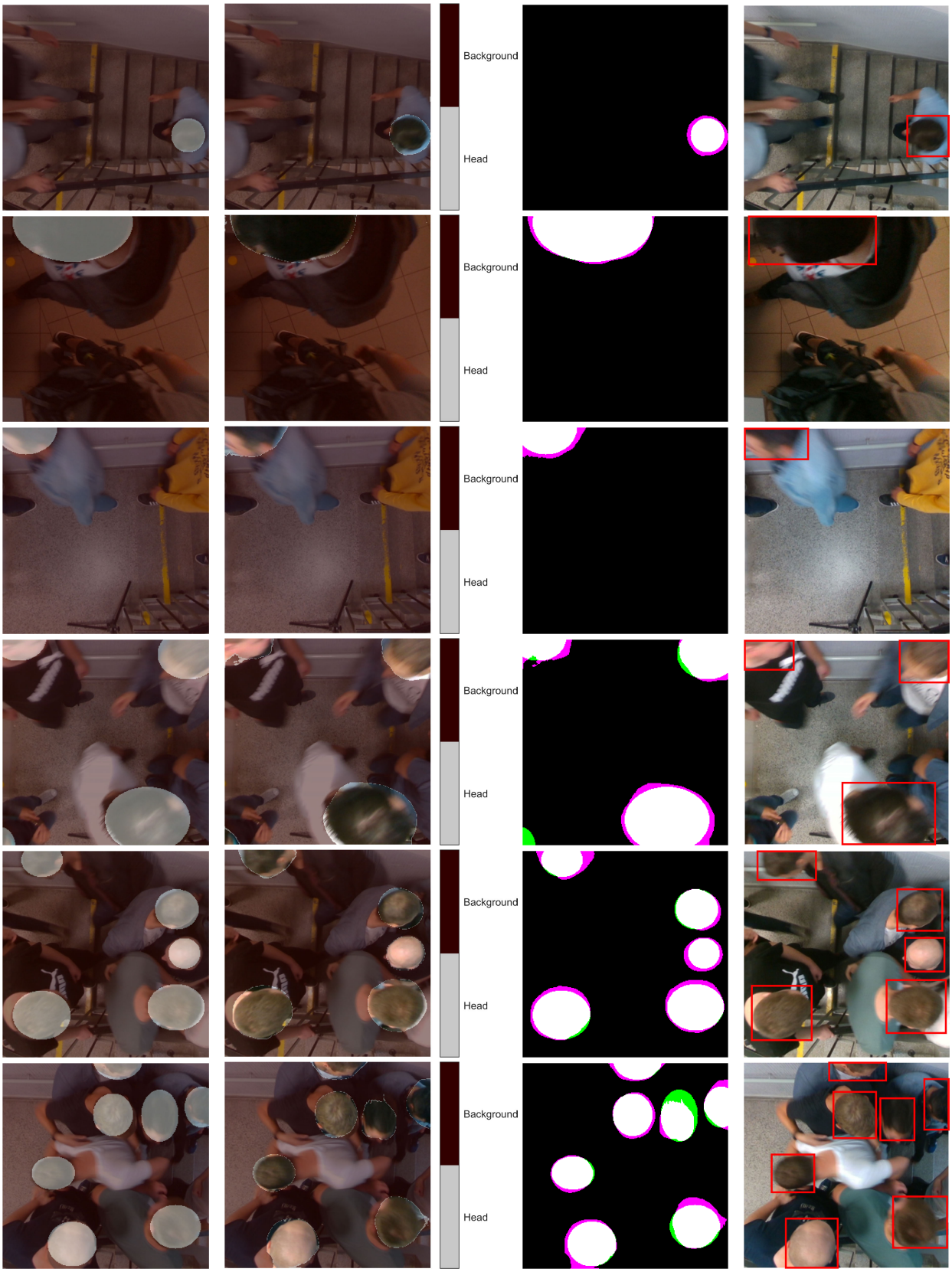


FIGURE 7. Experimental results on the first dataset. Columns (a), (b), (c), and (d) represent respectively: the tested image superimposed on its ground truth, image superimposed on its detected mask, ground truth on its corresponding detected mask, and the final object detection task.

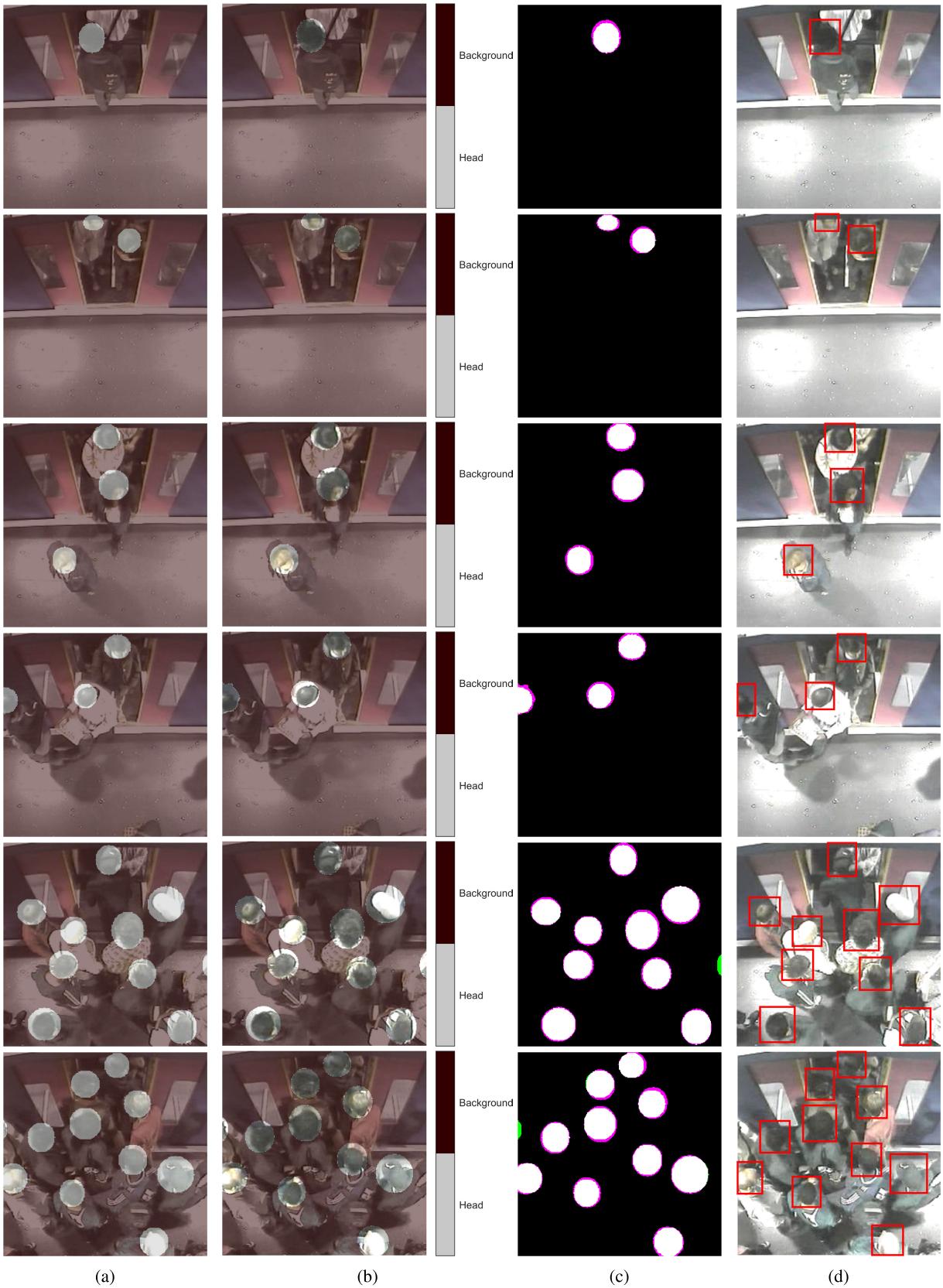


FIGURE 8. Experimental results on the second dataset. Columns (a), (b), (c), and (d) represent respectively: the tested image superimposed on its ground truth, image superimposed on its detected mask, ground truth on its corresponding detected mask and the final object detection task.

taken into account. In imbalanced class situations, true negative results often completely skew the perception of the performance. However, considering the BF score, a large number of true negatives leaves it unmoved.

Our method concentrates on reusing the relationship among the first network features, and broadcasting the important information by freezing them to the second network for further convolution operations. By concatenating extracted features with the original feature maps, the second network has more information about the input image, so it can compare and enhance the predicted boundary to match the ground truth boundary. This setting allows a considerable improvement in the efficiency of extraction and use of spatial features in CNNs without too many additional calculations. This explains the presented results.

In general, the number of trainable parameters, memory size, and response time are the important parameters to consider for implementation in real-time applications. Table 2 shows a comparison between these relevant features for the tested models. As we can observe, due to the low depth of the proposed network, the number of parameters that should be adjusted during the training period is much lower than the original U-Net, resulting in faster training and faster convergence.

The purpose of this network design was also to reduce the amount of memory required, and to increase speed, while at the same time, to increase the accuracy of the network behavior. Therefore, with a 49.61% reduction in the number of parameters and a very similar response time, compared to the U-Net network, our system has succeeded in achieving higher accuracy than other methods.

To investigate the advantage of the proposed system for small training data sets, we randomly selected a subset of 1%, 5%, 10%, 25%, 50% and 100% of the set of data and calculated the performance of the proposed system in terms of scores of monitored criteria versus the fraction of the amount of data used to train the models. We applied the same procedure for the U-Net network to compare the results.

Figure 6 demonstrates the strongest of the proposed systems in a small dataset for semantic segmentation. The first and second rows represent the result of the first and second datasets. The CSU-Net model performance is consistently better than U-Net by a solid margin. In this experiment, we showed that our strategy procedure helps the learning to be more effective in visual representations from a small number of images.

To show the overall performance of the proposed system, a comparative analysis was carried out on nine deep learning semantic segmentation models for two considered datasets. Table 3 shows the comparative results of our proposed system by comparing it with U-Net, Linknet, PSPNet, FPN, IRUNet, DR-Unet104, and CFPNet-M. Additionally, as a part of the ablation study, the proposed system is compared with the main block of the CSU-Net itself (referred to as SU-Net), and with the same architecture of two shallow U-shaped models, but without the concatenation between the main block and the enhancement block. This model is referred to as 2SU-Net. For qualitative evaluation of our method, we employed three widely used performance metrics for our image semantic segmentation, including the mean Accuracy, mean IoU, and the mean BF score. We have bolded the text that corresponds to the best results for all algorithms o visually facilitate the benchmarking of performance. The proposed system outperformed the other competitors in all cases, when considering mean Accuracy. The only competing model that provided a better result in some metrics is IRUNet. In those cases, however, the resulting metric values are close. Additionally, IRUNet provides significantly worse response times; see Table 2. Moreover, if CSU-Net, SU-Net and 2SU-Net are compared, the first one gives better results for all metrics. Therefore, a series of concatenation combinations between the main block and the enhancement block seems to provide a considerable advantage.

TABLE 3. Comparative analysis of our proposed model with four state-of-art models of deep learning semantic segmentation algorithms: U-Net, Linknet, PSPNet, FPN, IRUNet, DR-Unet104, and CFPNet-M. The bolded text corresponds to the best results.

		Mean Acc	Mean IoU	Mean BF Score
Private dat.	U-Net	91.12%	87.88%	81.23%
	CSU-Net	**96.66%**	**89.67%**	81.02%
	SU-Net	89.57%	85.57%	78.63%
	2SU-Net	92.92%	88.71%	77.50%
	Linknet	61.15%	49.14%	63.14%
	PSPNet	60.65%	49.45%	68.96%
	FPN	59.00%	49.19%	64.93%
	IRUNet	93.11%	89.35%	**84.14%**
	DR-Unet104	72.14%	64.19%	71.05%
	CFPNet-M	54.24%	50.26%	61.44%
UANDES	U-Net	97.18%	95.00%	97.82%
	CSU-Net	**97.79%**	95.25%	**98.40%**
	SU-Net	97.34%	95.04%	98.22%
	2SU-Net	97.38%	95.06%	98.31%
	Linknet	66.13%	48.81%	75.29%
	PSPNet	74.56%	48.75%	72.56%
	FPN	64.94%	48.81%	75.49%
	IRUNet	97.59%	**95.26%**	98.33%
	DR-Unet104	96.83%	93.38%	93.98%
	CFPNet-M	91.63%	90.29%	91.06%

Figures 7 and 8 show experimental tests for the 1st and 2nd datasets. The selected test images are of low to high complexity. The first column represents images labeled in pixels (ground truth) by superimposing them on their raw images. The second one represents the detected head mask superimposed on the RGB image. The third column represents a comparison of our system's results with the expected ground truth. The colors highlight areas where the segmentation results differ from the expected ground truth (the green color indicates false positive segmentation and the magenta color indicates true negative segmentation). The

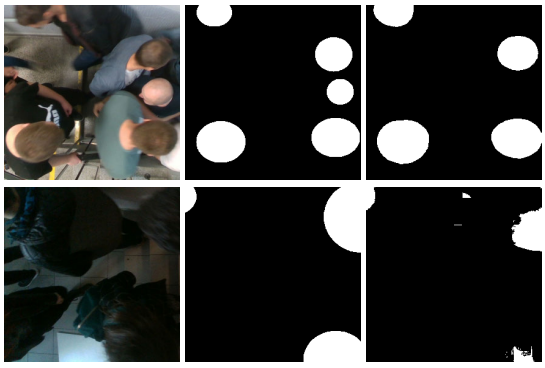


FIGURE 9. Examples of failure. The first column in each row represents the tested image, the second column is the ground truth, and the third column represents the response of CSU-Net. In the first row, the hairless head is not detected correctly, which is probably due to the under-representation of hairless persons in the training set. In the second row, the poorly detected head on the bottom right seems to blend in with the clothes of the other people in the image. In addition, the whole scene is poorly illuminated.

fourth column represents the final output of the system with all detected objects in red bounding boxes.

Experimental tests proved very satisfactory. We can see the strong effect by precisely detecting the heads that are very close to each other without overlapping them and considering them as a single head. Experimental tests also showed that the CSU-Net system detected almost all the heads correctly, even in a complex environment.

The evaluation of CSU-Net also aimed to capture significant failures of the presented model. Intuitively, the model may fail if there are objects in the scene that are similar to human heads and were not present in the training set. Other causes of failure can be poor lighting conditions, low-quality (blurry) image of the scene, or unexpected overlapping of people in the frame. Therefore, every response of CSU-Net was manually checked for failure. Several responses that can be classified as failures were found. Almost all of them were related to the private dataset, which can thus be described as more difficult to deal with. Surprisingly, all of them were false positives. Hence, no objects in the scene (bags, backpacks, suitcases) were falsely detected as heads. Typical examples of failures can be seen in Figure 9.

VI. CONCLUSION

Semantic segmentation suffers from many drawbacks, such as poorly predicted object contours. In this paper, our aim was to self-enhance the semantic segmentation result with minimal computational cost. For this purpose, we developed a new convolutional neural network architecture based on two shallow U-Net networks, which are strongly concatenated with each other. The goal was to force the second network to improve the output of the first block. The proposed strategy allowed the system to have the ability to learn additional convolutional mappings to reflect more information about the original input image when generating features of a higher level. Even though the proposed system has two encoder-decoder architectures, it has been designed with the consideration of reducing the amount of computational cost and memory requirement for training and testing. If we compare our proposed architecture with the original U-Net, the number of parameters in the proposed system is about 49.61% lower. A comparative analysis was carried out on several deep learning semantic segmentation state-of-art models for two datasets (one private and the other public). We also showed the advantage of the proposed system for small training data sets. This paper confirms that the idea of the concatenation of two networks, with reducing the trainable parameters, helps to improve the semantic segmentation and presents encouraging results. The results thus indicate that the proposed system can also be applied in several other semantic segmentation applications. Future work will be conducted on head object detection using semantic instance segmentation and head tracking on sequenced images.

REFERENCES

[1] X. Xiao, Y. Zhao, F. Zhang, B. Luo, L. Yu, B. Chen, and C. Yang, "BASeg: Boundary aware semantic segmentation for autonomous driving," *Neural Netw.*, vol. 157, pp. 460–470, Jan. 2023.

[2] R. Zhang, G. Li, T. Wunderlich, and L. Wang, "A survey on deep learning-based precise boundary recovery of semantic segmentation for images and point clouds," *Int. J. Appl. Earth Observ. Geoinf.*, vol. 102, Oct. 2021, Art. no. 102411.

[3] H. Abedi, S. Luo, V. Mazumdar, M. M. Y. R. Riad, and G. Shaker, "AI-powered in-vehicle passenger monitoring using low-cost mm-wave radar," *IEEE Access*, vol. 10, pp. 18998–19012, 2022.

[4] J. Wang, Y. Zhang, Y. Wei, Y. Hu, X. Piao, and B. Yin, "Metro passenger flow prediction via dynamic hypergraph convolution networks," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 12, pp. 7891–7903, Dec. 2021.

[5] J. H. Siregar and M. J. Budiman, "Optimal schedule in urban transportation to reduce the passenger crowded area," *Geographia Technica*, vol. 15, pp. 143–150, Apr. 2020.

[6] S. Chen, H. Fu, N. Wu, Y. Wang, and Y. Qiao, "Passenger-oriented traffic management integrating perimeter control and regional bus service frequency setting using 3D-pMFD," *Transp. Res. C, Emerg. Technol.*, vol. 135, Feb. 2022, Art. no. 103529.

[7] Y. Guo, Z. Liu, H. Luo, H. Pu, and J. Tan, "Multi-person multi-camera tracking for live stream videos based on improved motion model and matching cascade," *Neurocomputing*, vol. 492, pp. 561–571, Jul. 2022.

[8] Z. Qin, H. Liu, B. Song, M. Alazab, and P. M. Kumar, "Detecting and preventing criminal activities in shopping malls using massive video surveillance based on deep learning models," *Ann. Oper. Res.*, to be published, doi: 10.1007/s10479-021-04264-0.

[9] P. Skrabanek, P. Dolezel, Z. Nemec, and D. Stursa, "Person detection for an orthogonally placed monocular camera," *J. Adv. Transp.*, vol. 2020, pp. 1–13, Oct. 2020.

[10] M. Chouai, P. Dolezel, D. Stursa, and Z. Nemec, "New end-to-end strategy based on DeepLabv3+ semantic segmentation for human head detection," *Sensors*, vol. 21, no. 17, p. 5848, Aug. 2021.

[11] S. A. Velastin, R. Fernández, J. E. Espinosa, and A. Bay, "Detecting, tracking and counting people getting on/off a metropolitan train using a standard video camera," *Sensors*, vol. 20, no. 21, p. 6251, Nov. 2020.

[12] Y. Guo, Y. Liu, T. Georgiou, and M. S. Lew, "A review of semantic segmentation using deep neural networks," *Int. J. Multimedia Inf. Retr.*, vol. 7, no. 2, pp. 87–93, 2018.

[13] 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, Jan. 2017.

[14] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[15] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.* Munich, Germany: Springer, Oct. 2015, pp. 234–241, doi: 10.1007/978-3-319-24574-4_28.

[16] D. M. Vo and S.-W. Lee, "Semantic image segmentation using fully convolutional neural networks with multi-scale images and multi-scale dilated convolutions," *Multimedia Tools Appl.*, vol. 77, no. 14, pp. 18689–18707, Jul. 2018.

[17] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[18] 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.

[19] L. Hoyer, D. Dai, Y. Chen, A. Koring, S. Saha, and L. Van Gool, "Three ways to improve semantic segmentation with self-supervised depth estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11130–11140.

[20] X. Li, X. Li, L. Zhang, G. Cheng, J. Shi, Z. Lin, S. Tan, and Y. Tong, "Improving semantic segmentation via decoupled body and edge supervision," in *Proc. 16th Eur. Conf. Comput. Vis.* Glasgow, U.K.: Springer, 2020, pp. 435–452.

[21] H. Yin, C. Zhang, Y. Han, Y. Qian, T. Xu, Z. Zhang, and A. Kong, "Improved semantic segmentation method using edge features for winter wheat spatial distribution extraction from Gaofen-2 images," *J. Appl. Remote Sens.*, vol. 15, no. 2, May 2021, Art. no. 028501.

[22] F. H. Gil Zuluaga, F. Bardozzo, J. I. Rios Patino, and R. Tagliaferri, "Blind microscopy image denoising with a deep residual and multiscale encoder/decoder network," in *Proc. 43rd Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC)*, Nov. 2021, pp. 3483–3486.

[23] Y. Zhu, K. Sapra, F. A. Reda, K. J. Shih, S. Newsam, A. Tao, and B. Catanzaro, "Improving semantic segmentation via video propagation and label relaxation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8856–8865.

[24] A. Pasad, A. Gordon, T.-Y. Lin, and A. Angelova, "Improving semantic segmentation through spatio-temporal consistency learned from videos," 2020, *arXiv:2004.05324*.

[25] F. G. Zanjani and M. van Gerven, "Improving semantic video segmentation by dynamic scene integration," in *Proc. NCCV*, 2016, pp. 1–16.

[26] J. Colman, L. Zhang, W. Duan, and X. Ye, "DR-Unet104 for Multimodal MRI brain tumor segmentation," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Lecture Notes in Computer Science), vol. 12659. Cham, Switzerland: Springer, 2021, pp. 410–419, doi: 10.1007/978-3-030-72087-2_36.

[27] A. Lou, S. Guan, and M. Loew, "CFPNet-M: A light-weight encoder–decoder based network for multimodal biomedical image real-time segmentation," 2021, *arXiv:2105.04075*.

[28] A. Loukkal, V. Fremont, Y. Grandvalet, and Y. Li, "Improving semantic segmentation in urban scenes with a cartographic information," in *Proc. 15th Int. Conf. Control, Autom., Robot. Vis. (ICARCV)*, Nov. 2018, pp. 400–406.

[29] M. Divecha, *Improving Semantic Segmentation for Autonomous Vehicles Using Synthetic Images*. Merced, CA, USA: Univ. California, 2019.

[30] J. Huang, G. Liu, and B. Wang, "Semantic segmentation under a complex background for machine vision detection based on modified UPerNet with component analysis modules," *Math. Problems Eng.*, vol. 2020, pp. 1–13, Sep. 2020.

[31] Y.-H. Huang, J. Oramas, T. Tuytelaars, and L. V. Gool, "Do motion boundaries improve semantic segmentation?" in *Proc. WiML*, 2016, pp. 1–4.

[32] H. Wu, C. Liang, M. Liu, and Z. Wen, "Optimized HRNet for image semantic segmentation," *Expert Syst. Appl.*, vol. 174, Jul. 2021, Art. no. 114532.

[33] L. Zhang, H. Li, P. Shen, G. Zhu, J. Song, S. A. A. Shah, M. Bennamoun, and L. Zhang, "Improving semantic image segmentation with a probabilistic superpixel-based dense conditional random field," *IEEE Access*, vol. 6, pp. 15297–15310, 2018.

[34] B. Niu, "Semantic segmentation of remote sensing image based on convolutional neural network and mask generation," *Math. Problems Eng.*, vol. 2021, pp. 1–13, Jun. 2021.

[35] Y. He, B. Schiele, and M. Fritz, "Synthetic convolutional features for improved semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, Aug. 2020, pp. 320–336, doi: 10.1007/978-3-030-66823-5_19.

[36] C. Dong, "Image semantic segmentation method based on GAN network and ERFNet model," *J. Eng.*, vol. 2021, no. 4, pp. 189–200, Apr. 2021.

[37] Y. Zhu, Z. Zhang, C. Wu, Z. Zhang, T. He, H. Zhang, R. Manmatha, M. Li, and A. Smola, "Improving semantic segmentation via self-training," 2020, *arXiv:2004.14960*.

[38] D. Shen, Y. Ji, P. Li, Y. Wang, and D. Lin, "RANet: Region attention network for semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2020, pp. 1–12.

[39] H. Farsi and S. Mohammadzadeh, "Improvement in accuracy and speed of image semantic segmentation via convolution neural network encoder–decoder," *J. Inf. Syst. Telecommun.*, vol. 3, no. 23, p. 128, 2019.

[40] B. Wu, S. Zhao, W. Chu, Z. Yang, and D. Cai, "Improving semantic segmentation via dilated affinity," 2019, *arXiv:1907.07011*.

[41] A. Tran, A. Zonoozi, J. Varadarajan, and H. Kruppa, "PP-LinkNet: Improving semantic segmentation of high resolution satellite imagery with multi-stage training," in *Proc. 2nd Workshop Structuring Understand. Multimedia Heritage Contents*, Oct. 2020, pp. 57–64.

[42] D. Gritzner and J. Ostermann, "Using semantically paired images to improve domain adaptation for the semantic segmentation of aerial images," *ISPRS Ann. Photogramm., Remote Sens. Spatial Inf. Sci.*, vol. 2020, pp. 483–492, Aug. 2020.

[43] S. Jadon, "A survey of loss functions for semantic segmentation," in *Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol. (CIBCB)*, Oct. 2020, pp. 1–7.

[44] A. Chaurasia and E. Culurciello, "LinkNet: Exploiting encoder representations for efficient semantic segmentation," in *Proc. IEEE Vis. Commun. Image Process. (VCIP)*, Dec. 2017, pp. 1–4.

[45] (2018). *UNet*. [Online]. Available: https://github.com/zhixuhao/unet

[46] P. Yakubovskiy. (2019). *Segmentation Models*. [Online]. Available: https://github.com/qubvel/segmentation_models

[47] F. Gil. (2021). *Blind Microscopy Image Denoising With a Deep Residual and Multiscale Encoder/Decoder Network*. [Online]. Available: https://github.com/Fabio-Gil-Z/IRUNet

[48] J. Colman. (2021). *DR-UNet104*. [Online]. Available: https://github.com/jordan-colman/DR-Unet104

[49] A. Lou. (2021). *CFPNet-M: A Light-Weight Encoder–Decoder Based Network for Multimodal Biomedical Image Real-Time Segmentation*. [Online]. Available: https://github.com/AngeLouCN/CFPNet-Medicine

[50] *Intel Realsense Depth Camera D435*. Accessed: Oct. 15, 2022. [Online]. Available: https://www.intelrealsense.com/depth-camera-d435/

MOHAMED CHOUAI received the Ph.D. degree in signals, design of systems and their applications from the University of Mostaganem, Algeria, in 2020, in cooperation with the University of Cartagena, Spain, for the development of systems based on machine learning/deep learning technology and the image processing. He worked at the University of Pardubice, Czech Republic, as a Postdoctoral Researcher. He is currently a Researcher with the Alfred Wegener Institute, Germany.

PETR DOLEZEL (Member, IEEE) received the Ph.D. degree from the University of Pardubice, Czech Republic, in 2009. In 2017, he defended his habilitation thesis at Tomas Bata University. He is currently working as an Associate Professor and the Vice-Rector for research with the University of Pardubice. He is the author of more than 100 scientific contributions, including 20 journal articles and lectures at CORE-ranked conferences. His research interests include neural and evolutionary computation in process control and signal and image processing. In addition, he has been a leader or a member of research teams for a dozen research and development projects. As an academician, he led three Ph.D. students to a successful defense of their dissertations. He is a member of the Technical Program Committee of several international conferences and an active reviewer for numerous scientific journals. He intensively cooperates with research teams at the University of Burgos, Spain, and the Slovak University of Technology, Slovakia.