Ear semantic segmentation in natural images with Tversky loss function supported DeepLabv3+ convolutional neural network

Tolga Inan, Umit Kacar

***Abstract*—Semantic segmentation is a fundamental problem for computer vision. On the other hand, for studies in the field of biometrics, semantic segmentation is gaining more importance. Many successful biometric recognition systems require a high-performance semantic segmentation algorithm. This study presents an effective ear segmentation technique in natural images. A convolutional neural network is trained for pixel-based ear segmentation. DeepLab v3+ network structure, with ResNet-18 as the backbone and Tversky lost function layer as the last layer, has been trained with natural and uncontrolled images. The proposed network training is performed using only the 750 images in the Annotated Web Ears (AWE) training set. The corresponding tests are performed on the AWE Test Set, University of Ljubljana Test Set, and the Collection A of In-The-Wild dataset. For the Annotated Web Ears (AWE) dataset, intersection over union (IoU) is measured as 86.3% for the AWE database. To the best of our knowledge, this is the highest performance achieved among the algorithms tested on the AWE test set.**

***Index Terms*—Semantic Segmentation, Ear Segmentation, Biometrics, Convolutional Neural Networks, Tversky Loss Function.**

I. Introduction

BIOMETRIC systems have become an integral part of our lives for many years, and are now widely used by forensic, security, and law enforcement agencies. To use biometric systems more easily in daily life and achieve higher performance, researchers have focused on end-to-end and fully automated solutions. In particular, facial and ear biometrics come to the fore in the public sphere because the face and ear biometrics can be recorded in a non-cooperative manner.

As with the recognition studies in many biometric modalities, studies in ear recognition were first conducted with data recorded in a controlled environment. There are many ear recognition studies using ear datasets recorded in controlled environments[1], [2]. However, given the difficulties in practical applications, these studies lack applicability to real-world scenarios. In 2017, to overcome this drawback, the Unconstrained Ear Recognition Challenge (UERC) competition was organized by [3]. The UERC competition was held once again in 2019 [4], and the competitors were given an unconstrained ear dataset. This database is open to the public and available to the entire research community. Since the UERC competition's main focus was ear recognition, the ear images were manually cropped and given to the participants and the research community. With this aspect, the competition database was unsuitable for ear detection studies.

It is widely accepted that ear detection is the first and most crucial step in the ear biometric recognition line[5]. With the semantic segmentation method, the localization accuracy at the ear detection stage can be increased. This article provides an effective method for ear segmentation. Our contribution to this problem is twofold. The first is to choose a suitable model for training supported by the augmentation of the data and the selection of the model parameters. Our second contribution is to choose an appropriate loss function for the ear segmentation problem.

The rest of the paper is as follows: Chapter 2 summarizes relevant studies, Chapter 3 introduces the Proposed Method, Chapter 4 and Chapter 5 describe Experimental Setup and Experimental Results, respectively. The article ends with the Conclusion section in Chapter 6.

II. Related Work

Object detection is a fundamental task in computer vision, and the primary purpose of object detection is: "Which object is where?"[6]. In the object recognition problem, the solution algorithm tries to find the rectangle that will best encircle the object. In this type of research, marked data sets are required for the algorithms' training and testing phases. In these data sets, the positions of the objects are given by the rectangles surrounding them. On the other hand, semantic segmentation studies aim to classify the object-class at the pixel level. The semantic segmentation algorithm classifies pixels and labels pixels as the object-class or the background-class. In this study, the object-class is chosen as the human ear, and the scope is limited to ear semantic segmentation.

A. Convolution Neural Network based approaches for Semantic Segmentation

In semantic segmentation studies, convolutional neural network-based approaches stand out. In this context, the studies trying to solve the semantic segmentation problem with the convolutional neural network approach are briefly mentioned in this section. The study by [7] focuses on designing a deep

Tolga Inan is with the Department of Electrical - Electronics Engineering, Cankaya University, Ankara, 06790 TURKEY e-mail: tolga.inan@cankaya.edu.tr

Umit Kacar is with the Department of Electrical - Electronics Engineering, Cankaya University, Ankara, 06790 TURKEY e-mail: umitkacar@itu.edu.tr

Manuscript received Nov 15, 2021; accepted July 18, 2022.
DOI: 10.17694/bajece.1024073

neural network architecture with low latency operation for the semantic segmentation problem. They build a solution applicable to real-time applications. They report more than 10 fps image segmentation speed with an input image resolution of 640x360 for the practical road scene parsing datasets.

One of the earlier attempts to use convolutional neural networks (CNN) for semantic segmentation problem is the study of [8]. This study represents a fully deep CNN with the support of a probabilistic graphical model. The location-invariant nature of the features generated by convolutional neural networks may result in poor localization for semantic segmentation tasks. To overcome this bottleneck of poor localization, [8] employs a fully-connected conditional random field (CRF) as a probabilistic graphical model to extract spatial dependence in the semantic segmentation problem. The aforementioned study reports 71.6 % Intersection over Union (IoU) accuracy as the state-of-art, for the PASCAL VOC-2012 [9] semantic image segmentation test set.

[10] propose a solution called RefineNet, a multi-path refinement network, and they report an intersection-over-union score of 83.4 on the PASCAL VOC-2012 dataset. [11] propose a semantic segmentation framework with the capability of handling zero-labeled and few-labeled object classes, improving their approach which was reported in [12]. Their method employs indirect information acquired from semantic space via the semantic projection network. They report their results for the zero-label and few-label learning semantic segmentation experiments conducted on COCO-Stuff [13] and PASCAL VOC12 [9] datasets.

[14] propose a convolutional network for semantic segmentation with attention support. Their study facilitates a criss-cross network structure to capture horizontal and vertical contextual information around a particular pixel. The harvesting of contextual information is repeated so that each pixel can finally capture the dependencies from all pixels. They report the mean intersection over union (mIoU) 81.4 and 45.22 as scores on Cityscapes [15] test set and ADE20K [16] validation set, respectively. The proposed method is claimed to be both memory and computation effective for GPU implementation with the state of the art performance. In the study by Chen et al. [17], they use the dilated (atrous) convolution approach with an atrous spatial pyramid pooling module to extract multi-scale features. The proposed system, also known as, DeepLabv3 demonstrates competitive performance on PASCAL VOC12 [9] dataset.

The study by [18] proposes a task called panoptic segmentation to integrate the distinct tasks of semantic segmentation and instance segmentation. Hence, their study attempts to assign pixel-level labels and to detect each object instance simultaneously. To report panoptic segmentation performance, the authors represent a new metric named as panoptic segmentation metric. Panoptic segmentation metric is a hybrid metric that unifies the effects of the segmentation quality and the recognition quality.

The recent attempt by [19] focuses on the neural architecture search. The search for neural architecture is hierarchical and twofold: cell level search and network-level search. They avoid the hand-designing of the higher-level network structure. The problem formulation is continuous and allows the gradient-based architecture search. They report state-of-the-art performance on Cityscapes, PASCAL VOC12, and ADE20K datasets without any pretraining on ImageNet [20]. The study by [21] uses an adversarial training to learn from unlabeled data to achieve pixel-level classification. They report state-of-the-art performance for semi-supervised learning.

For further information on semantic segmentation, the reader is referred to detailed survey papers [22], [23], [6], [24], [25] on semantic segmentation using deep learning techniques.

B. Datasets for Ear Segmentation

Semantic segmentation approaches are usually supervised learning approaches; therefore, they require a sufficient amount of labeled data for the training stage and the test stage's performance measurement. In this section, four databases that are commonly used for ear segmentation studies are pointed-out.

[26] shared the Annotated Web Ears (AWE) dataset. This database consists of 1000 images (750 training images and 250 test images) of 100 subjects. They had collected the images from the web by a semi-automatic procedure. The largest image's size in the database is 473-by-1022 pixels, whereas that of the smallest image is 15-by-82 pixels. The average image size is reported to be 83-by-160 pixels. The binary masks indicating ear and non-ear classes are also distributed with the database.

[27] used 12500 images from the web, collected by the researchers and students from University of Ljubljana. The images are unconstrained, and they have different resolutions. They generate a semi-automatic procedure that obtains a pixel-wise class label (ear and non-ear) masks. They use the RefineNet [10] method to generate pixel-wise class label masks and manually examine the results for valid masks. Therefore they can conveniently extend the database size reported in [27]. At the date, the database have been obtained, there were 16765 pixel-wise class label masks available for the researchers.

[28] shared the UBEAR database. This database consists of 4412 gray-scale images and their pixel-wise class label masks. The images are recorded in dynamic lighting conditions with on-the-move subjects. The subjects did not pay attention to ear occlusions and their poses during the recording sessions.

[29] shares "In-the-Wild" database with images from collected Google Images. The images are gathered with ear-related tags, and the identities are not known. In this study, the Collection A set, in which the ear is manually annotated, is used . Fifty-five anatomically distinct landmark points around the ear regions (ascending helix, descending helix, helix, ear lobe, tragus, canal, antitragus, concha, etc.) are manually annotated. Collection A set includes 605 images, and the images are randomly divided into two disjoint sets of Collection A-training (500 images) and Collection A-testing (105 images). For the images shared by [29], 55 points are marked manually, but the masks for ear semantic segmentation are not available. These masks were obtained with the following procedure. A convex-hull is created from the 55 marked points. The points that determined the convex-hull are the outermost ones. The

outermost points helped to determine the outer border of the ear. The outer points were connected with the line segments, and the outer border was utterly determined. With the help of this determined closed exterior, the masks for semantic segmentation have been created.

C. Convolution Neural Network based approaches for Pixel-wise Ear Semantic Segmentation

In this section of the paper, four studies devoted specifically to the pixel-wise ear segmentation will be briefly outlined. This first one is by [30]. They propose a pixel-wise ear detection approach based on their convolution encoder-decoder network. Encoder–decoder architectures consist of two main stages. In the first stage, the input image is encoded to an abstract representation with the help of convolutional and pooling layers. Whereas in the second stage, the abstract representation is decoded into the desired output format. In [30], the detection pipeline has the assumption that there is a single face in the input image, and the aim is to detect at most two ears. The convolutional encoder-decoder is reported to perform well on image inputs that are recorded totally in unconstrained environments. They test the performance of their pixel-wise ear segmentation network on AWE dataset [26] and report the average accuracy, the average IoU, the average precision, the average recall as 99.4 %, 55.7%, 67.7% and 77.7%, respectively.

The second work is a recent paper by [27]. They propose a method called Mask R-CNN for the pixel-wise ear segmentation. The Mask R-CNN method consists of five stages: Convolutional backbone architecture, region proposal network, region of interest classifier, bounding box regressor, and detection pixel-wise masks. The performance of the method is tested on AWE dataset [26]. The performance indicators on the AWE dataset are as follows. The average IoU is 79.24%, the average precision is 92.04%, and the average recall is 84.14%.

Most of the biometric recognition studies include the detection and/or segmentation stage. The third study summarized as a pixel-wise ear segmentation approach given in [5] presents a complete ear recognition pipeline. [5] includes the pixel-wise ear segmentation stage and the ear segmentation is based on RefineNet [10]. RefineNet is an effective semantic segmentation network using residual convolution units, multi-resolution fusion, and chained residual pooling to achieve high-performance semantic segmentation. The ear segmentation performance is reported on the AWE dataset [26] and the corresponding performance figures for the average accuracy, the average IoU, the average precision, the average recall are 99.8 %, 84.8%, 91.7% and 91.6%, respectively.

The most recent study related to ear semantic segmentation is Context-aware Ear Detection Network (ContexedNet) [31]. ContexedNet has two stages. The first stage is a context-provider, and it generates the probability maps for facial regions. The second stage is the semantic segmentation stage, supported by the probability maps obtained in the first stage. The corresponding performance figures on the AWE dataset are 99.74 %, 81.46%, 89.07%, and 87.47%, for the average accuracy, the average IoU, the average precision, the average recall, respectively.

Although this study is about ear segmentation in pixel resolution, two other types of ear detection approaches are briefly mentioned in the following two sections. Firstly, the examples for the studies that detect curved boundaries or landmarks of the ear are given. Moreover, finally, the studies in the literature that detect the ear as a rectangular region are mentioned.

D. Studies those Detect Curved Boundaries or Landmarks of the Ear

[29] represents holistic and patch-based statistical deformable models to localize the ear landmarks. Thie holistic model is based on a shape, appearance, and deformation models. Inversion compositional algorithm, an efficient variant of gradient descent, is used for incremental wrapping. They also offer a patch-based active appearance model. Both holistic and patch-based models are tested on the 55-point ear landmark database. The 55-point ear landmark database has been prepared as a part of this study. Fitting accuracy for the landmark points is reported for the aforementioned database. [32] trains a small-sized CNN to localize 45 landmarks on the ear. They claim that the proposed method is robust to occlusion to some extent. [33] propose a method for finding physiological curves for the ear. The main focus of this study is finding interval curves of the ear (helix, antihelix, concha auriculae).

E. Studies those Detect the Rectangular Boundary for the Ear

The study of [34] offers an approach to detects ears from 2D images. Their method is based on the ensemble of convolutional neural networks (CNN). Three different CNNs had been trained for this purpose. In the next step, the weighted average of the outputs from three different CNNs was used, and ear detection was performed in this way. It has been reported that the performance achieved in this way is higher than the performance achieved with a single model.

In [35], a method based on Faster R-CNN is proposed for detecting ears in 2D images. They claim that they have improved the original R-CNN algorithm using multiple scales, resulting in faster and more accurate system performance.

Paper by [36] addresses the critical role of ear detectors that can work in unconstrained environments to have reliable biometric recognition systems. To tackle with ear detection problem in the wild, they offer two context-aware detection models based on CNNs. In this study, an accuracy of 99 % is reported at IOU 0.5 for the majority of the datasets.

The study proposed by [37] is an ear detection system based on CNN. Instead of employing a single CNN, they prefer to use three CNN, which are trained for different scales of ear images. The scales are obtained by cropping the ear region from the image with the small, medium, and large windows. The proposes are displayed to have higher performance than the Haar Cascade classifier. No other comparisons with recent methods are reported.

III. Proposed Method

A. Problem Definition

Semantic segmentation is basically the classification of images at the pixel level. A system that successfully performs semantic segmentation needs to assign the class label to each pixel correctly. There are many proven, cutting-edge models of semantic segmentation in the literature. Most of these models offer solutions for semantic segmentation problems for objects (trees, automobiles, animals, buildings, roads, etc.) that are frequently encountered in daily life. However, special adjustments and improvements are required to solve certain problems such as ear detection. These improvements can be expressed as follows:

- Obtaining the appropriate model for transfer learning
- Loss function selection

In this study, an efficient ear segmentation model was created by considering the items summarized above. Since it has been confirmed in the literature that transfer learning increases model accuracy in a shorter training period with fewer data, the transfer learning is used in model training. The following sections describe the improvements at each step.

B. Obtaining the appropriate model for transfer learning

DeepLab v3+ [38] is one of the state-of-art deep learning models for semantic image segmentation, where the goal is to assign semantic labels to every pixel in the input image. This study uses the DeepLab v3+ encoder-decoder network, and pre-trained model ResNet-18 [39] to make an effective start to the training process. ResNet-18 architecture requires less computing resources compared to ResNet-50. This advantage of the ResNet-18 architecture played an important role in the choice of this architecture and makes the overall system more efficient.

C. Loss function selection

The selection of the loss function is a fundamental issue having a significant effect on CNNs' performances. Many studies [40], [41], [42], [43], [44] have examined the importance of loss functions for semantic segmentation. Especially in Jadon's study [41], different error functions are discussed. They point out that the Tversky loss function generally generates optimal results. In this study, the Tversky loss function proposed by [45] is used in the last layer of our network. Tversky loss function is given Equation 1. In Equation 1, P and G are the sets of predicted and ground truth binary labels. On the other hand, α and β are the control parameters of the penalties for false positives and false negatives, respectively.

$$S(P,G;\alpha,\beta) = \frac{|PG|}{|PG| + \alpha|P\ G| + \beta|G\ P|} \tag{1}$$

Tversky loss function [45] is proposed for tackling the negative effects of the data imbalance in medical applications like lesion segmentation. In ear semantic segmentation, a similar problem arises. Only a small number of pixels belong to the ear class in the training and testing images. Considering this fact, the Tversky loss function is used to increase the ear segmentation network's overall performance.

IV. Experimental Setup

A. Data Augmentation

The performance of supervised training algorithms increases with the variety of training data. Data augmentation is used to increase the variety of training data to support our training process. Two ways of data augmentation are used. The first one is flipping the training image vertically at random (with 0.5 probability). Correspondingly, so that balanced training is carried out for human faces randomly oriented to the right and the left. The second way for data augmentation is changing the input image's scale randomly. The training images are scaled with a factor chosen randomly from [0.8, 1.2] closed interval. These two augmentation methods contribute to the increase of data diversity in the training set.

B. Evaluation Metric

Ear semantic segmentation is a pixel-based ear detection problem. Therefore, definitions for detection performance can be used in reporting semantic segmentation performance. True Positive (TP) refers to a pixel correctly labeled as the ear-class, while True Negative (TN) is used for pixels that are correctly labeled as the background. False Positive (FP) indicates a background pixel labeled as an ear. False Negative (FN) indicates an ear pixel labeled as the background. The other performance parameters are calculated in Equations 2-9 and the performance reporting is carried out with these calculated parameters based on these definitions of the TP, TN, FP, and FN.

$$Accuracy = 100 \times \frac{(TP+TN)}{(TP+TN+FN+FP)} \tag{2}$$

$$Precision = 100 \times \frac{TP}{(TP+FP)} \tag{3}$$

$$Recall = 100 \times \frac{TP}{(TP+FN)} \tag{4}$$

$$IoU = 100 \times \frac{TP}{(TP+FN+FP)} \tag{5}$$

$$F1\ Score = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{6}$$

$$FP\ Rate\ (FPR) = 100 \times \frac{FP}{(FP+TN)} \tag{7}$$

$$FN\ Rate\ (FNR) = 100 \times \frac{FN}{(FN+TP)} \tag{8}$$

$$E2 = \frac{FPR+FNR}{2} \tag{9}$$

C. Training Set

Increasing data diversity in semantic segmentation works helps to increase performance. However, to make a fair comparison with other ear semantic segmentation studies in the literature, the training set is limited to the training set (750 images) in the AWE dataset [26]. Throughout this study, only the training set in the AWE database was used for training purposes.

A label mask is also used for each training image. The label mask is a binary image showing labels (ear and background) for each pixel and has the same number of rows and columns as the training image.

D. Preliminary Experiments for Training Parameters Selection

The training performance of the CNNs is related to the initial conditions of the training process. It is not always possible to select all parameters optimally. In this study, some parameters are kept constant while some other parameters are chosen in a way that will increase the performance. In the following, the fixed parameters and the parameters selected to achieve high-performance objective are briefly described.

1) Fixed Parameters: The first fixed parameter is the size of the input image. The input image is set to 480 x 640 pixels. Another fixed parameter used is the mini-batch size. Due to the limit of the memory capacity of the GPU, the mini-batch value is set to 32.

2) Parameters selected to achieve high-performance objective: In this section, the process for selecting the parameters in order to design a high-performance ear semantic segmentation system is described.

Three parameters are defined as variables. The first of these parameters is the control parameters of the penalties for false positives parameter, that is α. Seven different values are considered for the α value. These values are 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, respectively.

The second parameter to be selected is the initial learn rate parameter. Three different values have been considered for the initial learn rate parameter. These values are 0.001, 0.0005 and 0.0001, respectively.

The third and last parameter to choose is the weight decay parameter. Two different values have been considered for the weight decay parameter. These values are 0.0005 and 0.0001, respectively.

3) Preliminary experiments to select parameters to achieve high-performance objective: Using the aforementioned parameter values, 42 (7x3x2) different parameter sets are determined.

For preliminary experiments, the preliminary-experiment-training-set and preliminary-experiment-test-set are defined. To make a fair comparison with the studies in the literature, only the AWE training set can be used at this step. So, first 500 images in the AWE training set are labeled as preliminary-experiment-training-set. The last 250 images in the AWE training set are labeled as preliminary-experiment-test-set. Therefore, only 750 images of AWE training set are used in preliminary experiments for parameter selection.

For preliminary experiments, the epoch number was kept constant as 2. Separate experiments were performed for 42 (7x3x2) parameter sets. Precision and accuracy graphs of these experiments are given in Figure 1. In Figure 1, axes represent α, initial learn rate and weight decay, respectively. Precision and accuracy values are expressed in the color codes given in the legend.

After evaluating the experimental results given in Figure 1, α was chosen as 0.55, initial learn rate as 0.001, and weight decay as 0.0001. Since α is set to 0.55; $\beta = (1 - \alpha)$, the control parameter of the penalties for false negatives, is calculated as 0.45. In this way, it is aimed to increase the system performance.

This approach has been used to make a more reasonable choice among the possible parameters. Therefore; we do not claim that the parameters chosen in this way are optimal. Care has been taken to use only the images in the AWE Training Set while making the parameter selection.

E. Test Sets

1) AWE Test Set: The test data were prepared by combining 250 images from the AWE data set [26]. Label masks given for these images were also used for performance calculations. Since this test set has been widely used in other studies in the field of ear semantic segmentation, using this test set helps us compare our ear semantic segmentation solution's performance realistically with other studies.

2) University of Ljubljana (UL) Test Set: This test set consists of 16765 images from the University of Ljubljana [27] dataset.

3) UBEAR Test Set: This test set is prepared by 4412 images from the UBEAR dataset [28].

4) Wild-1 Test Set: This test set consists of 500 training images of Collection A of "In-the-Wild" database gathered by [29].

5) Wild-2 Test Set: This test set consists of 105 testing images of Collection A of "In-the-Wild" database gathered by [29].

V. Experimental Results

A. Hardware and Software

The experiments were carried out on a personal computer with an AMD Ryzen 7 2700x processor and 32 GB of RAM. There is an Nvidia RTX 2080Ti GPU card in the setup. The operating system of the personal computer is Ubuntu 20.04 LTS. Training and tests are carried out in Matlab 2021a environment. The ADAM algorithm [46] is used as an optimizer.

B. Training with the Selected Parameters

The semantic segmentation network is trained with the parameters defined in Section IV-D. Training is performed for ten epochs with 750 images in the AWE training dataset [26]. The calculated loss values for ten epochs are shown in Figure 6.

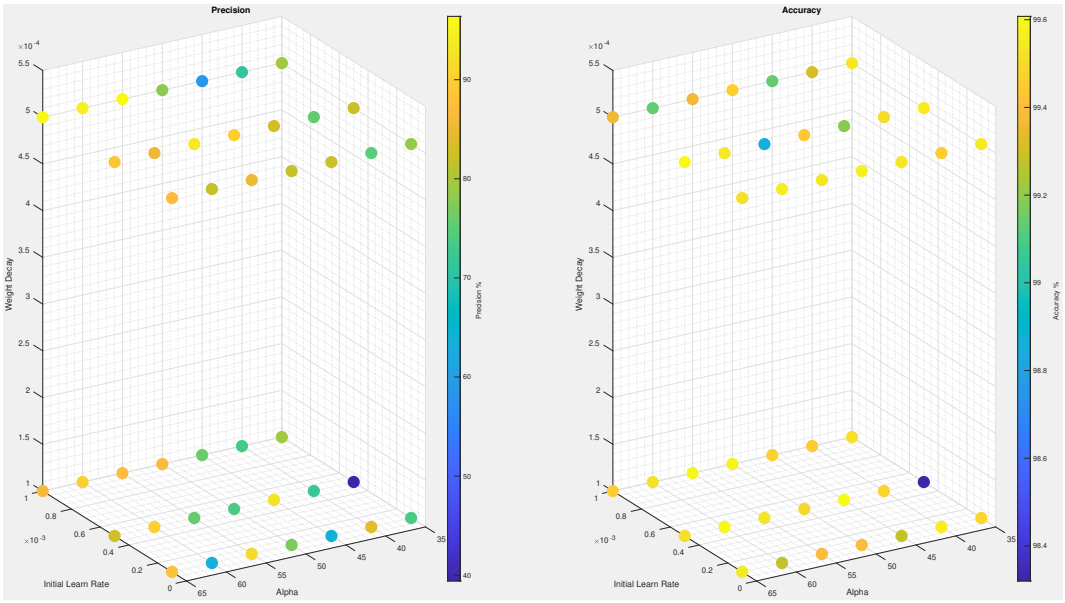

Fig. 1. Precision and Accuracy for the preliminary experiments on 250 images from AWE **Training** Dataset

C. Test Results for Ear Semantic Segmentation

1) AWE Test Set: The performance test has been carried on 250 test images on AWE dataset [26]. The performance results are summarised in Table I. The accuracy results are very high in all studies, including ours. This is basically due to the fact that the background-class labeled pixels are dominating in number in most of the images. A major performance metric in the semantic segmentation studies is the IoU. We report the IoU as 86.3%. For the AWE test set, the average precision is 93.5 %, and the average recall is 91.8 %. In Table I, high performance of our proposed algorithm on AWE dataset [26] is clearly observed. The average processing time for the images in the test set is 0.037 seconds.

The ear segmentation results are displayed for thirty-six test images having the highest IoU scores in the AWE test set Figure 2. Images are ordered with respect to IoU scores with row-major ordering. Thirty-six test images with the lowest IoU scores in the AWE test set are shown in Figure 3. The IoU scores are also represented as a histogram and are shown in Figure 4. It is observed from Figure 3 and Figure 4, only five images are having an IoU score lower than 0.8. Accuracy values for the images are given in Figure 5 as a histogram. We reported high-performance results for the AWE dataset. We also performed tests on other datasets to document the performance of our ear segmentation solution in different datasets. The results we obtained from these tests, together with the number of images in the datasets, are given in Table II. In the sections followed, we will briefly evaluate the results obtained in these datasets.

2) UL Test Set: The UL dataset [27] consists of images on the web collected by researchers and students from the University of Ljubljana. This dataset with 16765 images is quite large and suitable for performance measurement. The IoU value is reported on this dataset as 82.6 % and the precision value as 93.5 % (Table II). Performance values in this extensive database of images recorded in an uncontrolled environment support the high performance of the proposed ear segmentation method.

3) UBEAR Test Set: The UBEAR dataset [28] consists of moving subjects, gray-scale images collected under varying lighting conditions. In addition, the subjects did not pay attention to their pose and ear occlusion. Due to the movement of the subjects and the variable lighting, degradation in sharpness in some images are present. With all these aspects, the UBEAR dataset has the most challenging conditions among the datasets used in this study. The IoU value is reported on this dataset as 55.9 % and the precision value as 85.3 % (Table II). Performance values obtained in this dataset are lower compared to the performance figures in other datasets. The decrease in performance is considered to be due to grey-scale images, moving and non-cooperative subjects, and variable lighting conditions.

4) Wild-1 Test Set: Wild-1 Test Set [29] consists of 500 images. The IoU value is reported on this dataset as 77.1 % and the precision value as 80.2. % (Table II). Images of the "In-the-Wild" dataset are collected from Google Images, and the subjects are not cooperative. This independent test set confirms the high values obtained for the performance metrics of the proposed ear segmentation network.

5) Wild-2 Test Set: Wild-2 Test Set [29] consists of 105 images. The IoU value is reported on this dataset as 79.7 % and the precision value as 81.8. % (Table II).

D. Ablation Study

In this study, the DeepLabv3+ convolutional neural network's last layer is updated using Tversky Loss Layer. An ablation study has been conducted to demonstrate the positive contribution of this update. For this purpose, the training is performed with the AWE Training Set using the Cross-Entropy Loss Layer as the last layer of the DeepLabv3+ convolutional

Fig. 2. Our ear segmentation results for the test images in the AWE dataset (with highest IoU scores).

neural network. The cross-entropy loss layer is the original loss layer in the DeepLabv3+ structure. Tests are performed on test sets using the convolutional neural network trained using the cross-entropy loss layer. All other hyper-parameters of the CNN are kept constant; therefore, the effect of the loss layer on the overall performance is displayed. The results with cross-entropy loss are given in Table III. These results confirm the positive impact of the Tversky Loss Layer on the overall performance.

TABLE I
COMPARISON OF TEST RESULTS FOR EAR SEGMENTATION ON AWE DATASET

Method	Accuracy	IoU	Precision	Recall	E2
Ped-Ced-Alt[30]	99.2	50.8	62.5	78.5	24.6
Ped-Ced[30]	99.4	55.7	67.7	77.7	22.2
Mask R-CNN[27]	—	79.24	92.04	84.14	—
RefineNet[5]	99.8	84.8	91.7	91.6	7.6
ContexedNet[31]	99.74	81.46	89.07	87.47	—
Our Study	**99.8**	**86.3**	**93.5**	**91.8**	**4.1**

TABLE II
OUR TEST RESULTS WITH TVERSKY LOSS FOR EAR SEGMENTATION ON DIFFERENT TEST SETS

Dataset	Images	Accuracy	IoU	Precision	Recall	E2
AWE	250	99.8	86.3	93.5	91.8	4.1
UL	16765	99.8	82.6	93.5	87.6	6.2
UBEAR	4412	99.3	55.9	85.3	61.8	19.2
WILD-1	500	99.8	77.1	80.2	95.2	2.5
WILD-2	105	99.8	79.7	81.8	96.8	1.7

TABLE III
OUR TEST RESULTS WITH CROSS ENTROPY LOSS FOR EAR SEGMENTATION ON DIFFERENT TEST SETS

Dataset	Images	Accuracy	IoU	Precision	Recall	E2
AWE	250	99.76	79.13	93.89	83.42	8.32
UL	16765	99.77	73.99	94.39	77.40	11.32
UBEAR	4412	99.24	45.92	91.07	48.09	25.99
WILD-1	500	99.75	73.62	83.83	85.81	7.16
WILD-2	105	99.76	77.48	83.94	90.97	4.60

E. Limitations of the Study

This study uses challenging databases for ear semantic segmentation, and high-performance results are reported for ear segmentation. The subjects in these databases are not cooperative and look in various directions. Again, in these databases, there are ear images with different scales. With these aspects, the results we obtained for ear-segmentation are enlightening about the performance of the proposed method. The study's main limitation is that the segmentation data used is specific to ear-segmentation only. Therefore, it is impossible to comment on the validity and performance of the proposed method in general semantic segmentation problems. Examining the proposed method in different segmentation problems and data types will help make more general evaluations.

VI. CONCLUSION

This article proposed a semantic segmentation method to segment the ears in the natural images recorded in uncontrolled environments. Our method is supported by selecting

Fig. 3. Our ear segmentation results for the test images in the AWE dataset (with lowest IoU scores).

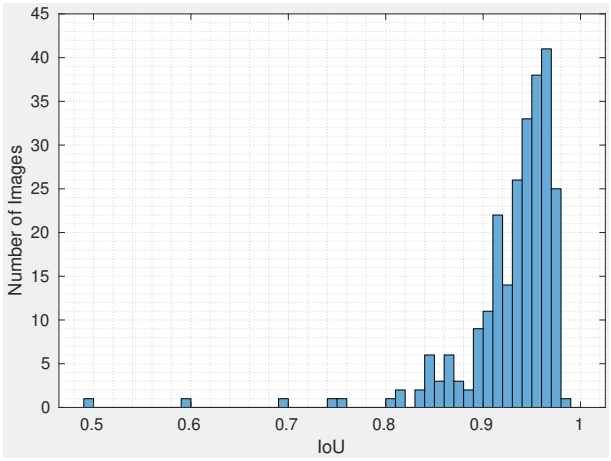

Fig. 4. Histogram for Intersection-over-Union(IoU) for AWE test set [26].

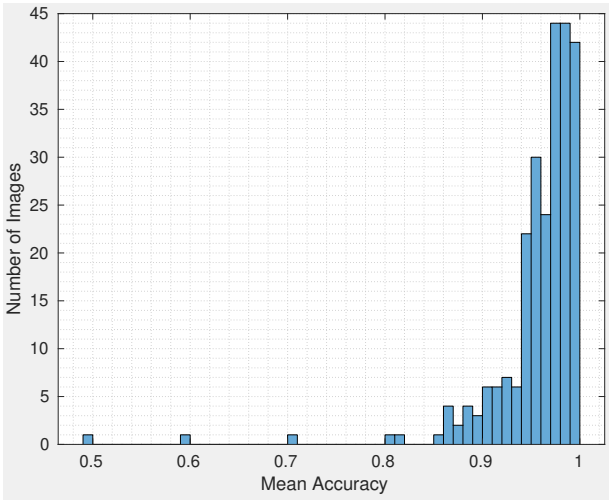

Fig. 5. Histogram for Accuracy for AWE test set [26].

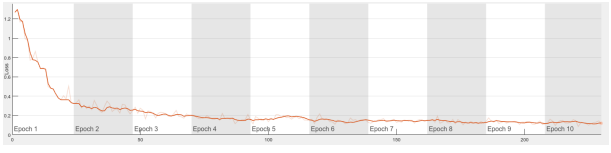

Fig. 6. Loss vs Iterations for Training on AWE Training Set

the appropriate model for transfer learning, and the appropriate loss function. AWE training set is used as the training data. DeepLab v3+ [38] encoder-decoder network and pre-trained model ResNet-18 [39] are used in this study to implement an accurate and effective ear semantic segmentation solution. The Tversky loss function [45] has been used in the final stage of our network to overcome the negative effects of the unbalanced distribution of the ear and background classes. The proposed algorithm's performance has been tested, and to the best of our knowledge, we have reported the highest scores related to performance on the AWE dataset [26]. We have also tested ear segmentation system performance on different challenging datasets and obtained test results displaying high accuracy and high precision of the proposed ear segmentation system.

One of our authors' recent study [47] is about high-performance ear recognition systems. In our future work, we will use the high performance ear semantic segmentation solution we developed in this publication, together with our knowledge of ear recognition to implement a very high performance, end-to-end, fully automatic ear recognition system. In our future studies, we plan to achieve successful ear recognition performance on the images recorded in natural and uncontrolled environments.

As another future work we plan to explore the few-label semantic segmentation. In this study, all of the images in the training set were completely labeled. Recently, there have been efforts to achieve semantic segmentation with few labels [11]. We also aim to extend our solution within this direction so that it will be possible to complete the training process with less labeled data.

ACKNOWLEDGMENT

Authors would like to thank Dr. Ziga Emeršic, Dr. Vitomir Struc, and their research team for sharing the ear databases, which made this study possible.

REFERENCES

[1] A. Abaza, A. Ross, C. Hebert, M. A. F. Harrison, and M. S. Nixon, "A survey on ear biometrics," *ACM Computing Surveys*, vol. 45, no. 2, pp. 1–35, Feb. 2013, number: 2 Reporter: ACM Computing Surveys. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2431211.2431221

[2] A. Pflug and C. Busch, "Ear biometrics: a survey of detection, feature extraction and recognition methods," *IET Biometrics*, vol. 1, no. 2, pp. 114–129, Jun. 2012, number: 2 Reporter: IET Biometrics. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/iet-bmt.2011.0003

[3] Z. Emersic, D. Stepec, V. Struc, P. Peer, A. George, A. Ahmad, E. Omar, T. E. Boult, R. Safdaii, Y. Zhou, S. Zafeiriou, D. Yaman, F. I. Eyiokur, and H. K. Ekenel, "The unconstrained ear recognition challenge," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*, Oct. 2017, pp. 715–724, meeting Name: 2017 IEEE International Joint Conference on Biometrics (IJCB) Reporter: 2017 IEEE International Joint Conference on Biometrics (IJCB) ISSN: 2474-9699.

[4] Z. Emersic, A. K. S. V, B. S. Harish, W. Gutfeter, J. N. Khiarak, A. Pacut, E. Hansley, M. P. Segundo, S. Sarkar, H. J. Park, G. P. Nam, I.-J. Kim, S. G. Sangodkar, U. Kacar, M. Kirci, L. Yuan, J. Yuan, H. Zhao, F. Lu, J. Mao, X. Zhang, D. Yaman, F. I. Eyiokur, K. B. Özler, H. K. Ekenel, D. P. Chowdhury, S. Bakshi, P. K. Sa, B. Majhi, P. Peer, and V. Štruc, "The Unconstrained Ear Recognition Challenge 2019," in *2019 International Conference on Biometrics (ICB)*, 2019, pp. 1–15.

[5] Z. Emersic, J. Krizaj, V. Struc, and P. Peer, "Deep Ear Recognition Pipeline," in *Recent Advances in Computer Vision: Theories and Applications*, ser. Studies in Computational Intelligence, M. Hassaballah and K. M. Hosny, Eds. Cham: Springer International Publishing, 2019, pp. 333–362, reporter: Recent Advances in Computer Vision: Theories and Applications. [Online]. Available: https://doi.org/10.1007/978-3-030-03000-1_14

[6] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object Detection in 20 Years: A Survey," *arXiv:1905.05055 [cs]*, May 2019, reporter: arXiv:1905.05055 [cs] arXiv: 1905.05055. [Online]. Available: http://arxiv.org/abs/1905.05055

[7] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A Deep Neural Network Architecture for Real-Time Semantic Segmentation," *arXiv:1606.02147 [cs]*, Jun. 2016, reporter: arXiv:1606.02147 [cs] arXiv: 1606.02147. [Online]. Available: http://arxiv.org/abs/1606.02147

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs," *arXiv:1412.7062 [cs]*, Jun. 2016, reporter: arXiv:1412.7062 [cs] arXiv: 1412.7062. [Online]. Available: http://arxiv.org/abs/1412.7062

[9] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, Jun. 2010.

[10] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path Refinement Networks for High-Resolution Semantic Segmentation," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, HI: IEEE, Jul. 2017, pp. 5168–5177, meeting Name: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Reporter: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). [Online]. Available: http://ieeexplore.ieee.org/document/8100032/

[11] Y. Xian, S. Choudhury, Y. He, B. Schiele, and Z. Akata, "Semantic Projection Network for Zero- and Few-Label Semantic Segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 8248–8257, meeting Name: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Reporter: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [Online]. Available: https://ieeexplore.ieee.org/document/8953827/

[12] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-Shot Learning - A Comprehensive Evaluation of the Good, the Bad and the Ugly," *arXiv:1707.00600 [cs]*, Aug. 2018, arXiv: 1707.00600. [Online]. Available: http://arxiv.org/abs/1707.00600

[13] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1209–1218.

[14] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "CCNet: Criss-Cross Attention for Semantic Segmentation," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. Seoul, Korea (South): IEEE, Oct. 2019, pp. 603–612, meeting Name: 2019 IEEE/CVF International Conference on Computer Vision (ICCV) Reporter: 2019 IEEE/CVF International Conference on Computer Vision (ICCV). [Online]. Available: https://ieeexplore.ieee.org/document/9009011/

[15] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, "The cityscapes dataset for semantic urban scene understanding," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 3213–3223.

[16] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[17] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking Atrous Convolution for Semantic Image Segmentation," *arXiv:1706.05587 [cs]*, Dec. 2017, reporter: arXiv:1706.05587 [cs] arXiv: 1706.05587. [Online]. Available: http://arxiv.org/abs/1706.05587

[18] A. Kirillov, K. He, R. Girshick, C. Rother, and P. Dollar, "Panoptic Segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 9396–9405, meeting Name: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Reporter: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [Online]. Available: https://ieeexplore.ieee.org/document/8953237/

[19] C. Liu, L.-C. Chen, F. Schroff, H. Adam, W. Hua, A. L. Yuille, and L. Fei-Fei, "Auto-DeepLab: Hierarchical Neural Architecture Search for Semantic Image Segmentation," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. Long Beach, CA, USA: IEEE, Jun. 2019, pp. 82–92, meeting Name: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Reporter: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). [Online]. Available: https://ieeexplore.ieee.org/document/8954247/

[20] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.

[21] S. Mittal, M. Tatarchenko, and T. Brox, "Semi-Supervised Semantic Segmentation with High- and Low-level Consistency," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019, reporter: IEEE Transactions on Pattern Analysis and Machine Intelligence. [Online]. Available: https://ieeexplore.ieee.org/document/8935407/

[22] A. Garcia-Garcia, S. Orts-Escolano, S. Oprea, V. Villena-Martinez, P. Martinez-Gonzalez, and J. Garcia-Rodriguez, "A survey on deep learning techniques for image and video semantic segmentation," *Applied Soft Computing*, vol. 70, pp. 41–65, Sep. 2018, reporter: Applied Soft Computing. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S1568494618302813

[23] S. Minaee, Y. Boykov, F. Porikli, A. Plaza, N. Kehtarnavaz, and D. Terzopoulos, "Image Segmentation Using Deep Learning: A Survey," *arXiv:2001.05566 [cs]*, Jan. 2020, reporter: arXiv:2001.05566 [cs] arXiv: 2001.05566. [Online]. Available: http://arxiv.org/abs/2001.05566

[24] F. Lateef and Y. Ruichek, "Survey on semantic segmentation using deep learning techniques," *Neurocomputing*, vol. 338, pp. 321–348, Apr. 2019, reporter: Neurocomputing. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S092523121930181X

[25] I. Ulku and E. Akagunduz, "A Survey on Deep Learning-based Architectures for Semantic Segmentation on 2D images," *IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING*, p. 14, 2019, reporter: IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING.

[26] Z. Emersic, V. Struc, and P. Peer, "Ear recognition: More than a survey," *Neurocomputing*, vol. 255, pp. 26–39, Sep. 2017, reporter: Neurocomputing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S092523121730543X

[27] M. Bizjak, P. Peer, and Z. Emersic, "Mask R-CNN for Ear Detection," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. Opatija, Croatia: IEEE, May 2019, pp. 1624–1628, meeting Name: 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO) Reporter: 2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO). [Online]. Available: https://ieeexplore.ieee.org/document/8756760/

[28] R. Raposo, E. Hoyle, A. Peixinho, and H. Proença, "UBEAR: A dataset of ear images captured on-the-move in uncontrolled conditions," *2011 IEEE Workshop on Computational Intelligence in Biometrics and Identity Management (CIBIM)*, pp. 84–90, 2011.

[29] Y. Zhou and S. Zaferiou, "Deformable Models of Ears in-the-Wild for Alignment and Recognition," in *2017 12th IEEE International Conference on Automatic Face Gesture Recognition (FG 2017)*, 2017, pp. 626–633.

[30] Z. Emersic, L. L. Gabriel, V. Struc, and P. Peer, "Convolutional encoder–decoder networks for pixel-wise ear detection and segmentation," *IET Biometrics*, vol. 7, no. 3, pp. 175–184, May 2018, number: 3 Reporter: IET Biometrics. [Online]. Available: https://digital-library.theiet.org/content/journals/10.1049/iet-bmt.2017.0240

[31] Z. Emersic, D. Susanj, B. Meden, P. Peer, and V. Struc, "Contexednet: Context–aware ear detection in unconstrained settings," *IEEE Access*, vol. 9, pp. 145 175–145 190, 2021.

[32] C. Cintas, C. Delrieux, P. Navarro, M. Quinto-Sánchez, B. Pazos, and R. Gonzalez-José, "Automatic Ear Detection and Segmentation over Partially Occluded Profile Face Images," *Journal of Computer Science and Technology*, vol. 19, no. 01, p. e08, Apr. 2019, number: 01 Reporter: Journal of Computer Science and Technology. [Online]. Available: http://journal.info.unlp.edu.ar/JCST/article/view/1097

[33] X. Zhang, L. Yuan, and J. Huang, "Physiological Curves Extraction of Human Ear Based on Improved YOLACT," in *2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT*, 2020, pp. 390–394.

[34] I. I. Ganapathi, S. Prakash, I. R. Dave, and S. Bakshi, "Unconstrained ear detection using ensemble-based convolutional neural network model: Unconstrained ear detection using ensemble-based convolutional neural network model," *Concurrency and Computation: Practice and Experience*, p. e5197, Feb. 2019, reporter: Concurrency and Computation: Practice and Experience. [Online]. Available: http://doi.wiley.com/10.1002/cpe.5197

[35] Y. Zhang and Z. Mu, "Ear Detection under Uncontrolled Conditions with Multiple Scale Faster Region-Based Convolutional Neural Networks," *Symmetry*, vol. 9, no. 4, p. 53, Apr. 2017, number: 4 Reporter: Symmetry. [Online]. Available: http://www.mdpi.com/2073-8994/9/4/53

[36] A. Kamboj, R. Rani, A. Nigam, and R. Jha, "CED-Net: context-aware ear detection network for unconstrained images," *Pattern Analysis and Applications*, 2020.

[37] W. Raveane, P. L. Galdámez, and M. A. González Arrieta, "Ear Detection and Localization with Convolutional Neural Networks in Natural Images and Videos," *Processes*, vol. 7, no. 7, p. 457, Jul. 2019, number: 7 Reporter: Processes. [Online]. Available: https://www.mdpi.com/2227-9717/7/7/457

[38] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, vol. 11211, pp. 833–851, reporter: Computer Vision – ECCV 2018. [Online]. Available: http://link.springer.com/10.1007/978-3-030-01234-2_49

[39] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[40] N. Abraham and N. M. Khan, "A novel focal tversky loss function with improved attention u-net for lesion segmentation," in *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 2019, pp. 683–687.

[41] S. Jadon, "A survey of loss functions for semantic segmentation," in *2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, 2020, pp. 1–7.

[42] W. Yuan and W. Xu, "Neighborloss: a loss function considering spatial correlation for semantic segmentation of remote sensing image," *IEEE Access*, vol. 9, pp. 75 641–75 649, 2021.

[43] D. Duque-Arias, S. Velasco-Forero, J.-E. Deschaud, F. Goulette, A. Serna, E. Decencière, and B. Marcotegui, "On power jaccard losses for semantic segmentation," in *VISAPP 2021: 16th International Conference on Computer Vision Theory and Applications*, 2021.

[44] H. Harkat, J. M. P. Nascimento, A. Bernardino, and H. F. Thariq Ahmed, "Assessing the impact of the loss function and encoder architecture for fire aerial images segmentation using deeplabv3+," *Remote Sensing*, vol. 14, no. 9, 2022. [Online]. Available: https://www.mdpi.com/2072-4292/14/9/2023

[45] S. S. M. Salehi, D. Erdogmus, and A. Gholipour, "Tversky loss function for image segmentation using 3D fully convolutional deep networks," *arXiv:1706.05721 [cs]*, Jun. 2017, reporter: arXiv:1706.05721 [cs] arXiv: 1706.05721. [Online]. Available: http://arxiv.org/abs/1706.05721

[46] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," *arXiv:1412.6980 [cs]*, Jan. 2017, arXiv: 1412.6980. [Online]. Available: http://arxiv.org/abs/1412.6980

[47] U. Kacar and M. Kirci, "ScoreNet: Deep cascade score level fusion for unconstrained ear recognition," *IET Biometrics*, vol. 8, no. 2, pp. 109–120, 2018, number: 2 Publisher: IET.

BIOGRAPHIES

Tolga Inan received the B.Sc., M.Sc., and Ph.D.degrees in electrical and electronics engineering from the Middle East Technical University, Ankara, Turkey, in 2000, 2003, and 2011, respectively. He is currently with the Department of Electrical-Electronics Engineering, Cankaya University, Ankara. His areas of research include machine learning, biometric recognition, computer vision, and power quality analysis.

Umit Kacar completed his Master's degree in Electronics Engineering from Istanbul Technical University in 2013 and his PhD in Electronics Engineering from Istanbul Technical University in 2019. He is currently working as a Senior Machine Learning Engineer in a private company. His research interests are biometrics, fusion, ensemble, artificial intelligence, machine learning and computer vision.