Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

ContexedNet: Context–aware Ear Detection in Unconstrained Settings

ŽIGA EMERŠIČ[1], DIEGO SUŠANJ[2], BLAŽ MEDEN[1], PETER PEER[1] (SENIOR MEMBER, IEEE), AND VITOMIR ŠTRUC[3] (SENIOR MEMBER, IEEE)

[1]Faculty of Computer and Information Science, University of Ljubljana, SI-EU)
[2]Faculty of Engineering, University of Rijeka, HR-EU
[3]Faculty of Electrical Engineering, University of Ljubljana, SI-EU

Corresponding author: Žiga Emeršič (e-mail: ziga.emersic@fri.uni-lj.si) and Diego Sušanj (e-mail: dsusanj@riteh.hr).

This research was supported in parts by the ARRS Research Programs P2–0250 (B) "Metrology and Biometric Systems" and P2–0214 (A) "Computer Vision".

ABSTRACT Ear detection represents one of the key components of contemporary ear recognition systems. While significant progress has been made in the area of ear detection over recent years, most of the improvements are direct results of advances in the field of visual object detection. Only a limited number of techniques presented in the literature are domain–specific and designed explicitly with ear detection in mind. In this paper, we aim to address this gap and present a novel detection approach that does not rely only on general ear (object) appearance, but also exploits contextual information, i.e., face–part locations, to ensure accurate and robust ear detection with images captured in a wide variety of imaging conditions. The proposed approach is based on a *Contex*t–aware *E*ar *D*etection *Net*work (ContexedNet) and poses ear detection as a semantic image segmentation problem. ContexedNet consists of two processing paths: *i) a context–provider* that extracts probability maps corresponding to the locations of facial parts from the input image, and *ii) a dedicated ear segmentation model* that integrates the computed probability maps into a context–aware segmentation-based ear detection procedure. ContexedNet is evaluated in rigorous experiments on the AWE and UBEAR datasets and shown to ensure competitive performance when evaluated against state–of–the–art ear detection models from the literature. Additionally, because the proposed contextualization is model agnostic, it can also be utilized with other ear detection techniques to improve performance.

INDEX TERMS ear detection, ear biometrics, biometrics, deep learning

I. INTRODUCTION

Ear detection is a crucial component and typically the first step in modern ear recognition systems. Poorly designed ear detection models adversely affect the performance of all downstream tasks of the recognition system, including normalization procedures, feature extraction techniques and classification approaches. Designing efficient and robust ear detection techniques is, therefore, critical for the overall performance of biometric ear recognition systems, as also emphasized by visible research in this area [1]–[4].

Recent work on ear detection focuses mainly on deep learning models and in particular on convolutional neural networks (CNNs). At the coarsest level this work can be partitioned into two main groups: *i) detection techniques* [5]–[7] and *ii) segmentation approaches* [1], [8]. Detection techniques build on advances in the area of visual object detection and include techniques designed around recent detection frameworks, such as region proposal CNNs (R-CNNs) [9], [10], masked region proposals CNNs (Masked R-CNNs) [11] and related models [12]–[14]. Segmentation-based methods, on the other hand, approach ear detection as a segmentation problem and exploit advances made in the area of semantic image segmentation [15]–[17]. Both detection and segmentation–based solutions have been shown to ensure competitive performance for ear detection on a wide variety of datasets and imaging conditions [1], [6], [7]. However, most of the techniques presented in the literature so far are generic and not designed specifically for ear detection. In other words, existing models exploit visual ear appearances for the detections/segmentation procedure, but treat ears as any other objects in the process. No specific information unique to the problem of ear detection is typically utilized,

leading to suboptimal detection performance.

To address this gap, we present in this paper a novel approach to ear detection that in addition to ear appearance also relies on contextual information to boost performance. Specifically, the proposed approach models the anatomy of the human head and incorporates information about the location of facial parts into the ear detection procedure. As a result, additional constraints are taken into account during the detection/segmentation step, which contributes towards improved performance. The detection framework, called *Contex*t-aware *E*ar *D*etection *Net*work (ContexedNet), falls into the group of segmentation–based approaches discussed above and exhibits the following characteristics:

- *Pixel-level detection*: Competing detection models typically return only a bounding box of the ear region and often assume that a single ear is present in the image [6], [7]. ContexedNet, on the other hand, produces pixel–level segmentation masks of an arbitrary number of ears and, hence, is more general and works under minimal assumptions.
- *Specificity and robustness:* ContexedNet is conditioned on information about face–part locations and is, therefore, designed specifically for the problem of ear detection - not general object detection. As demonstrated in the experimental section, the proposed model also ensures better robustness to challenging imaging conditions, which makes it applicable in ear recognition systems operating in unconstrained settings.
- *Modularity:* ContexedNet consists of two main components: *i) a context–provider* that extracts information on facial part locations from the given input images, and *ii) a segmentation model* that integrates the extracted information into a context–aware detection procedure. In this work, both components are implemented with recent CNN models from the literature. However, the proposed contextualization is *model agnostic* and can be implemented with any model with suitable characteristics. ContexedNet can, therefore, be expected to further improve with future advancements in either face–part detection or semantic image segmentation.

To demonstrate the applicability of ContexedNet for ear detection[1], experiments are conducted on the AWE [1] and UBEAR [18] datasets and comparisons with competing methods from the literature are presented. Experimental results show that ContexedNet achieves state–of–the–art performance on all experimental datasets, but also that the proposed contextualization is beneficial and helps to improve the performance of different baseline (segmentation) models.

In summary, the main contributions of this paper are:

- A novel framework for ear detection, called ContexedNet, that incorporates contextual information into the detection procedure by modeling human head anatomy and (implicitly) constrains ear detection results to the vicinity of predefined facial parts.
- A model contextualization procedure that forms the basis for ContexedNet and can be used in related problem domains and with different base/backbone models.
- A comprehensive experimental assessment and analysis of the proposed framework and contextualization procedure as well as a rigorous comparative evaluation with existing state-of-the-art techniques. To ensure reproducibility of the reported results, all code and models are made publicly available[2].

The rest of the paper is structured as follows: In Section II relevant prior work is discussed. In Section III ContexedNet is introduced and its main characteristics are elaborated on. The experimental evaluation of the proposed detection model is presented in Section IV. The paper concludes with a summary of the main findings and directions for future work in Section VI.

II. RELATED WORK

A considerable amount of prior work addressed the problem of ear detection, as summarized by recent surveys on this topic [2], [3], [19]. This prior work can in general be divided into three main groups: *i) image–processing techniques*, *ii) learning–based methods*, and *iii) deep–learning models*. Details on the three groups are given below.

A. IMAGE–PROCESSING TECHNIQUES

Techniques from this group rely on the low–level image–processing operations that try to highlight edge information, identify shapes or match ear characteristics to predefined ear templates in either the original pixel domain or some transformed space [20]–[23]. A common characteristic of this group of techniques is that they are computationally simple, rely on relatively strong assumptions (e.g., presence of one ear, full profile image input, etc.) and often degrade in performance when applied in challenging imaging conditions, where large variations in ear appearances can be expected.

Arbab–Zavar and Nixon [20], for example used the Hough transform to identify elliptically shaped regions that correspond to ears in the input images. A conceptually similar approach was later also described by Prajwal et al. in [21]. In [22], [23], the Canny edge detector was used to extract edges from ear images and the curves corresponding to the outer helix of the ears were used as features to identify ear regions in images. An approach based on the distance transform and template matching was introduced by Prakash et al. [24]. The same authors also proposed solutions that analyzed graphs constructed from an edge

[1]Note that the term *detection* is used in this paper to refer to the detection of the region–of–interest (ROI) in the ear image and corresponds to a segmentation task when used in the context of ContexedNet. We note that in the computer vision literature the term is typically used to describe bounding box detection tasks.

[2]http://awe.fri.uni-lj.si/ [After review!]

map of the ear image [25], [26] and an approach relying on skin–color filtering [27]. In [28], a detection technique based on the image ray transform was proposed. The transform first highlights the tubular structures of the ear and later exploits the highlighted structures for ear detection. Relevant techniques from this group also include [29], [30].

As can be seen from the above discussion, early ear detection techniques tried to model visual ear characteristics explicitly and use the modeled characteristics for the detection procedure. The approach proposed in this work is similar to the surveyed techniques in that it also tries to exploit visual ear characteristics for detection, but instead of using hand–crafted approaches to do so, it learns relevant characteristics for ear detection directly from the training data, leading to better overall detection performance.

B. LEARNING–BASED METHODS

The second group of techniques relies on learning–based methods for ear detection. Techniques from this group treat ear detection as a classification problem, where image patches sampled from the input images are typically classified into one of two classes: ears and others objects. Learning–based methods represent an evolution of image–processing based techniques that shifted in focus from designing descriptive features to designing efficient classification models for ear detection. Techniques from this group typically result in better performance than image–processing methods and are capable of handling a wider range of appearance variability, but require a considerable amount of data for training [31], [32].

Islam et al. [33] proposed an AdaBoost–based approach to ear detection that falls into this group of methods. The approach, inspired by the seminal Viola–Jones algorithm [34], relies on low–level Haar features for image (or patch) representation and a cascaded Adaboost classifier for the detection. An improved version of the approach was later presented by Abaza et al. in [35] and also by Liu and Liu in [36] where a skin color model was incorporated into the detection procedure, to further improve performance. A variation of the same idea was also discussed in [37].

Our detection approach is similar conceptually to learning–based models in that it also aims to learn a classifier (though at the pixel–level) that is capable of identifying image pixels that belong to ear regions. However, it relies on a more recent class of machine learning models (i.e., CNNs) that are able to exploit more descriptive image features (and not only low–level texture descriptors) and consequently handle a wider range of image variability.

C. DEEP–LEARNING MODELS

Most recent ear detection techniques from the literature rely on deep learning. While in essence, this group is also learning–based, the main difference with the group, discussed in the previous section, is in the way the detection problem is approached. While learning–based methods use a separate stage for feature extraction (or data representation)

(a) Spatial context (b) Face–part context

FIGURE 1: Standard object detection frameworks model (spatial) context through a multi-scale analysis, as shown on the left. ContexedNet uses a different strategy and exploits information about face part locations to model context, as illustrated on the right. Note how the face parts precondition the location of the ear region. The figure is illustrative and best viewed in color.

and patch classification, and typically utilize manually engineered or hand–crafted features for detection, deep learning models jointly learn image features as well as a classifier for detection in an (usually) end–to–end manner.

Zhang and Mu [7], for example, proposed an ear detection approach based on Faster Region-based Convolutional Neural Networks (Faster R-CNNs). The model built on advances in the domain of general object detection and was shown to ensure highly competitive results on the UBEAR [18] and UND dataset (J2 Collection) [38]. Another conceptually similar approach was later presented by El–Naggar et al. in [39] and again demonstrated the power of the Faster R-CNN framework for ear detection.

Tomczyk and Szczepaniak [40] presented a solution for ear detection based on geometric deep learning. The proposed model allows for the application of CNNs on graphs and defines convolutional filters with the use of Gaussian mixture models (GMMs). Based on this concept, the authors design a competitive detection framework that exhibits considerable robustness to rotations (i.e., it is rotation equivariant) as well as other desirable characteristics.

Raveane et al. [41] described a CNN-based approach to ear detection that utilizes a multi–path model topology and detection grouping to identify ear regions in the images. The main idea behind this approach is to look for ears at multiple scales akin to the contextual modules used in modern object detection frameworks, such as [42], [43], with the goal of improving detection performance. A similar idea was also explored by Kamboj et al. in [6], which applied generic object detection models with contextual modules for the task of ear detection. These works are related to the approach proposed in this paper in that they also exploit contextual information (multi–scale view of ears), but they rely on conceptually different approaches within standard detection frameworks. CentexedNet, on the other hand, builds on

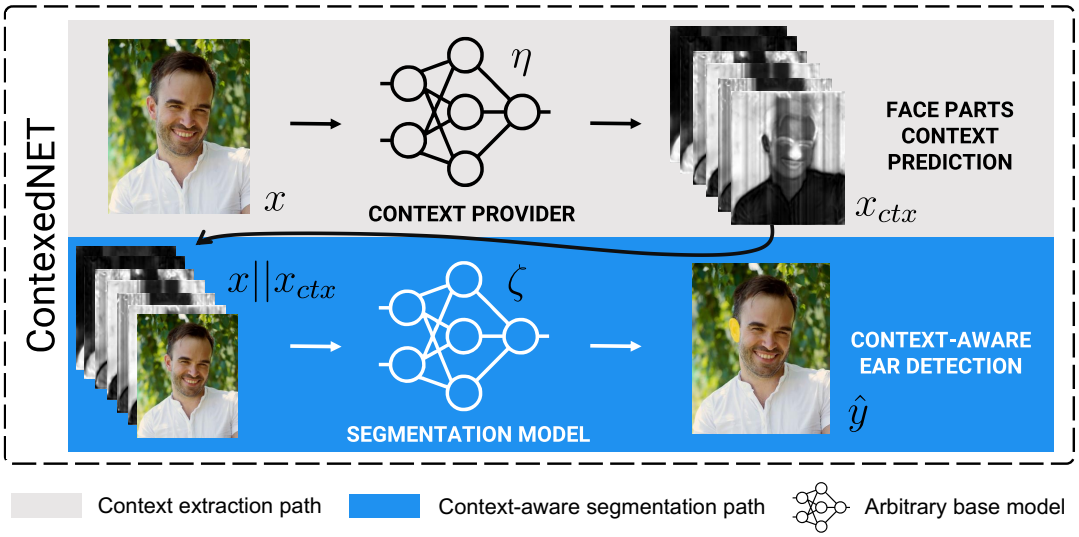


FIGURE 2: High-level overview of the ContexedNet ear detection framework. ContexedNet represents a two-path deep learning framework, where the first path (shown at the top) extracts contextual information in the form of feature maps encoding facial-part locations, and the second path (shown at the bottom) uses these feature maps jointly with the input image for segmentation of the ear region. The framework is *model agnostic* and can be implemented with any base/backbone model in either of the two processing paths. The main novelty of the framework comes from the contextualization procedure that infuses cues on face-part locations into the segmentation procedure and, therefore, has a strong geometric motivation.

advances in semantic segmentation and relies (for the most part) on a different type of context, defined by face part locations.

Specifically, ContexedNet extends our previous work on segmentation–based ear detection with PED–CED [1] to also consider high–level contextual information in addition to the raw input image. While in [1], an auto-encoder like model was used and a single image served as the input for segmenting the ear region, CentexedNet improves on this framework by also incorporating predictions about the head anatomy into the segmentation procedure. As we show in the experimental section, such an approach leads to highly competitive segmentation/detection results and reduces semantically unreasonable errors, where ears are detected in the image background or other body parts.

III. CONTEXT-AWARE EAR DETECTION

Using contextual information to improve the performance of various vision tasks has a rich history in computer vision [44]–[46] and has led to successful applications in object recognition, tracking [47], [48], biometrics [49]–[51], video analytics [52], surveillance and security [53] and even affective computing [54]. In the object detection literature, contextual information is commonly accounted for through a multi-scale analysis, where objects of interest are examined at different scales, as illustrated in Figure 1(a)[3]. This type of approach allows modern detection models to learn not only from object appearances but to also consider contextual information, i.e., from the surroundings of the object. For ContexedNet, described in this section, we consider a different approach and do not utilize only such standard spatial context. Instead, we propose to incorporate cues on face part locations into the detection procedure. Such cues have a geometrical motivation, as illustrated in Figure 1(b), and provide strong priors on the location of ears in the images. We note st this point that the main contribution of this paper is not in a new network or model architecture, but in the overall framework that infuses contextual information on face-part locations into the ear detection/ssegmentation procedure. As already emphasized in the introductory section, the framework itself is model agnostic and can be used with any recent backbone segmentation model. Details on ContexedNet are given in the following sections.

[3]The image shown was taken from the Flickr page of Maria Rantanen and was modified from its original appearance. The image is distributed under the Creative Commons license.

A. OVERVIEW OF CONTEXEDNET

A high-level overview of ContexedNet is presented in Figure 2. The model consists of two distinct processing paths: *(i) a context provider* that extracts feature maps encoding information on face-part locations, and *(ii) a dedicated segmentation model* that takes both, the raw input image as well as the generated feature maps as input and predicts a segmentation mask corresponding to the ear region(s).

Formally, the model can be described as follows. Given an input RGB image $x \in \mathbb{R}^{w \times h \times 3}$ from some training set $\mathcal{X}$ with corresponding segmentation targets $y \in \mathbb{R}^{w \times h}$, where $\mathcal{X} = \{(x_i, y_i)\}_{i=1}^{N}$ and N is the number of training

FIGURE 3: Examples of the intermediate feature representations, x_{ctx}, that encode face–part locations. Shown are feature maps for (from left to right): the skin, the eyebrows, the nose, the mouth, and the neck. The probability output of the face parser is used as the feature representation to ensure that face–part locations are not encoded in binary form. Shown are 5 out of the c_f generated representations.

examples[4], the goal of ContexedNet is to learn a mapping ψ parameterized by θ_ψ, such that the predicted output

$$\hat{y} = \psi(x; \theta_\psi) \in \mathbb{R}^{w \times h}, \tag{1}$$

is as close to the ground truth y as possible for every sample in $\mathcal{X}$. ContexedNet achieves this by first modeling constellations of face parts with an auxiliary context-provider η that generates an intermediate representation x_{ctg} from x, i.e.,

$$x_{ctx} = \eta^{(l)}(x; \theta_\eta) \in \mathbb{R}^{w \times h \times c_f}, \tag{2}$$

where c_f is the number of feature maps and the superscript l indicates the x_{ctx} is derived from the l-th layer of η. Next, it feeds the generated representations together with the input image to the segmentation network ζ that then produces the final segmentation result, i.e.:

$$\hat{y} = \zeta(x||x_{ctx}; \theta_\zeta), \tag{3}$$

where $||$ denotes the concatenation operator and $\theta_\psi = [\theta_\eta, \theta_\zeta]$. The main components and outputs generated within ContexedNet are marked in Figure 2. Details on the two processing paths of ContexedNet are described in the following sections.

B. THE CONTEXT PROVIDER

To extract information on face–part locations from the input image x, the context provider is designed around a face parser η that generates a parsing map $p \in \mathbb{R}^{w \times h \times c_f}$ from x with c_f segmented facial components. While any face parser can be utilized for this purpose, we select DeepLabV3+ [17] as the base model for our implementation due to its state–of–the–art performance and the fact that an open source implementation is readily available. The model is trained independently of the segmentation path of ContexedNet using a standard binary cross–entropy loss for each facial component, i.e. [55]–[57],

$$\mathcal{L}_{bce}^{(cp)}(p, \hat{p}; \theta_\eta) = -\sum_{i=1}^{c_f} p_i \log(\hat{p}_i) + (1 - p_i) \log(1 - \hat{p}_i), \tag{4}$$

[4]Note that we drop the sample subscript i in the following discussion to keep the notation uncluttered.

where p_i stands for the i-th facial part (i.e., the i-th channel) of the ground truth parsing map p, $\hat{p}_i$ denotes the corresponding prediction, and the superscript cp denotes the fact that the loss is associated with the *context provider* of ContexedNet. The number of facial parts c_f is an open hyper–parameter of the context provider and depends on the annotations present in the training data.

The parsing map p generated by η consists of c_f binary (face–parts) masks. To avoid a binary encoding of face–part locations and ensure consistent (i.e., intensity) inputs for the segmentation path of ContexedNet, the *probability output* of the context provider for each of the c_f channels is used as the intermediate feature representation x_{ctx} of the face parts. A few illustrative examples of the feature maps (for the neck, the eyebrows, the nose, the mouth and the neck) generated with the presented procedure are shown in Figure 3.

C. CONTEXT–AWARE SEGMENTATION NETWORK

Once the feature representations x_{ctx} are generated, they are fed as an additional input to the segmentation path of ContexedNet. Here, the feature representations are concatenated with the original RGB image x and used to constrain the ear detection/segmentation model, so it generates semantically reasonable predictions and avoids erroneous results, where segmentation masks are predicted in image areas without the correct context. The segmentation path is trained based on concatenated inputs $x_{con} = x||x_{ctx} \in \mathbb{R}^{w \times h \times (c_f + 3)}$ again using a standard binary cross–entropy loss, i.e. [55], [59]:

$$\mathcal{L}_{bce}^{(sp)}(y, \hat{y}; \theta_\zeta) = -\sum_{i=1}^{c_f} y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}), \tag{5}$$

where y and $\hat{y}$ are the ground truth ear segmentation mask and the corresponding model prediction, respectively. The superscript sp indicates that the loss is associated with the segmentation path of ContexedNet. Once the model is trained, ear segmentation masks are generated in accordance with Eq. (3).

For the implementation of the segmentation path, we again use a DeepLabV3+ model and explore different backbones for its implementation. However, note that in general the outlined context–aware segmentation procedure is model agnostic, so any segmentation model could be used for the implementation. Nonetheless, DeepLabV3+ was selected as the backbone for our experiments because: (i) source code for the model is publicly available (important for reproducibility), (ii) it ensures state-of-the-art results for a wide variety of segmentation tasks [17], and (iii) the fact that the model heavily relies on atrous convolutions that help to capture spatial context similarly to context modules typically used with contemporary detection models.

D. TRAINING PROCEDURE AND DEPLOYMENT

ContexedNet is trained using a two-stage procedure. In the first stage, we learn to predict c_f representations that encode

	CelebAMask-HQ [58]	AWE [1]	UBEAR [18]
Number of images	$30,000$	$1,000$	$4,412$
Number of classes	19	2	2
Resolution	512×512	Various	1280×960
Class list	19 classes corresponding to face parts and accessories, such as skin, nose, eyeglasses, eyes, mouth, upper lip, lower lip, hair, etc.	Background, ear	Background, ear
Used for	Training of context provider ($30,000$ images)	Training and validation (750 images); Testing (250 images)	Training and validation (2206 images); Testing (2206 images)

TABLE 1: Properties of the three datasets used in the experiments. CelebAMask-HQ is utilized for training the context provider, AWE for training and testing of the segmentation model, and UBEAR for testing only. The experimental protocol is provided in the bottom row of the table.

(a) AWE samples (b) UBEAR samples

FIGURE 4: Samples images and corresponding pixel–level ground truth masks from: (a) the AWE-W dataset, and (b) the UBEAR 1.0 dataset. Note that the images feature in these datasets were not collected in constrained conditions, as this is the case with many existing ear datasets. As result, the images exhibit considerable appearance variability that makes them challenging for ear detection/segmentation.

face-part locations by minimizing the training objective from Eq. (4) over a datasets with suitable ground truth annotations. This training step optimizes the parameters θ_η of the face parser η. In the second stage, we learn to predict the final segmentation masks based on the input image x and the extracted contextual information x_{ctx} by minimizing the loss from Eq. (5). This second stage results in optimized parameters θ_ζ for the context-aware segmentation model ζ. Once the two models are learnt, the final segmentation mask $\hat{y}$ corresponding to the ear region in the image is generated based on Eq. (3).

IV. EXPERIMENTAL SETUP

Several experiments were designed to evaluate the performance of the proposed ContexedNet. A summary of the setup used for these experiments is presented in the reminder of this section.

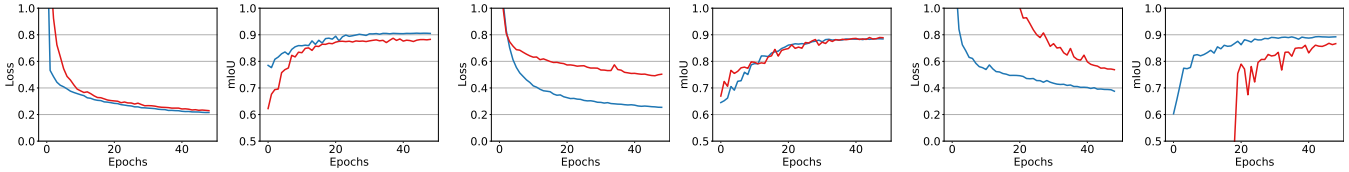
A. DATASETS AND EXPERIMENTAL SPLITS

Three datasets were selected for the experimental evaluation: CelebAMask-HQ [58], Annotated Web Ears (AWE) [1], and UBEAR 1.0 [18]. A high-level overview of the datasets and the experimental protocol used is provided in Table 1.

The first experimental dataset, CelebAMask-HQ, contains $30,000$ images of size 512×512 pixels with pixel–level annotations of 19 face components and accessories. Images in this dataset were collected from the web and feature a wide range of appearance variability. CelebAMask-HQ is used to train the context provider of ContexedNet.

The second dataset, AWE, consists of 1000 ear images of 100 subjects, captured in unconstrained conditions, as illustrated in Figure 4(a). Images in this datasets were again collected from the web and come with pixel–level annotations of the ear region. Because the acquisition conditions vary from image to image, the AWE data exhibits variability across environments (outdoor vs. indoor), illumination conditions, occlusions, image quality, but also demographic factors, such as age, gender and ethnicity. These characteristics make it highly challenging for the task of ear detection/segmentation. Images from the AWE dataset are used to train (750 images) and test (250 images) the segmentation model of ContexedNet, with the train and test split being subject and image disjoint.

The last dataset used in the experiments is UBEAR. This dataset was captured in an indoor environment under room lighting, but in an uncooperative scenario, where the subjects did not pose in perfect profile view during data acquisition. The UBEAR images, therefore, vary in terms of pose, blur and overall image quality, as shown in Figure 4(b). Similarly to AWE, UBEAR also comes with pixel–level annotations (i.e., binary masks) of the ear region. UBEAR is used in the experiments for the performance evaluation to demonstrate how ContexedNet generalizes to other data characteristics and to compare the performance of the proposed framework to standard bounding-box based



(a) ResNet loss. (b) ResNet mIoU. (c) MobileNet loss (d) MobileNet mIoU (e) Xception loss (f) Xception mIoU

FIGURE 5: Comparison of the training characteristics for the three backbone models, ResNet [60], MobileNet [61] and Xception [62] with (in blue) and without (in red) contextual information. Results are presented in terms of the (training) cross–entropy loss and the mIoU on the validation data. Note how the addition of contextual information helps with the convergence of the segmentation model both in terms of pace as well all as performance reached. Best viewed in color.

TABLE 2: Computational complexity of the training procedure when learning the three backbone models of ContexedNet's segmentation path with and without context. Run–time complexity is also given. Results are reported for the experimental hardware used, i.e., an NVIDIA GeForce Titan XP with 12GiB of VRAM. The symbols h, m and s stand for hours, minutes and seconds, respectively.

Segmentation Backbone	Setting	Training time		Test time†
		Context provider	Segmentation model	
ResNet [60]	w/o Ctx.	n/a	$2h\ 45m$	$\approx 0.05s$
	w Ctx.	$\approx$ 1 day	$2h\ 35m$	$\approx 0.11s$
MobileNet [61]	w/o Ctx.	n/a	$45m$	$\approx 0.03s$
	w Ctx.	$\approx$ 1 day	$35m$	$\approx 0.09s$
Xception [62]	w/o Ctx.	n/a	$2h\ 45m$	$\approx 0.06s$
	w Ctx.	$\approx$ 1 day	$2h\ 35m$	$\approx 0.12s$

† The test time complexity is given per sample and was computed as an average over 100 test images.

ear detectors..

B. PERFORMANCE MEASURES

Results are reported using two performance measures in order to facilitate comparisons with previously published works, i.e., overall segmentation accuracy (Acc) and mean intersection over union (mIoU). Accuracy is typically defined in the ear–detection literature as the ratio between the number of correct detections and the overall number of annotated ear areas. However, the criterion for deciding on correct or incorrect predictions varies in the literature. Here, we use the definition from [1], where accuracy is defined through a segmentation tasks and consider both the number of correctly classified ear pixels as well as the number of correctly classified non–ear pixels, averaged over all n test images, i.e. [63], [64]:

$$Acc = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i + TN_i}{d_i}, \tag{6}$$

where d_i denotes the number of pixels, TP_i stands for the number of true positives, i.e., the number of pixels correctly classified as part of the ear, TN_i stands for the number of true negatives, i.e., the number of pixels correctly classified as non-ear pixels, in the i-th image. However, because this measure is not weighted by the representation of classes (i.e., the ground truth number of ear and non–ear pixels), it is impacted most by the majority class. i.e., the background.

We, therefore, also report the mean intersection over union (IoU) for the experiments, which is defined as follows [65], [66]:

$$mIoU = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FP_i + FN_i}, \tag{7}$$

where n again denotes the number of test images, and FP_i and FN_i denote the number of false positives (i.e., ear pixels classified as non-ear pixels) and the number of false negatives (i.e., non-ear pixels classified as ear pixels), for the i-th test image, respectively. A value of 1 means that the detected and annotated ear areas overlap perfectly, while a value of 0 indicates a completely failed detection, i.e. no detection at all or a detection outside the actual ear area.

Additionally, we also report precision, recall and F1 scores for the *ear segmentation* task in order to provide better overall understanding of the performance of our models and to compare it more easily with other works from the literature. Here, precision, recall and F1 are defined as follows [66], [67]:

$$Precision = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FP_i}, \tag{8}$$

$$Recall = \frac{1}{n}\sum_{i=1}^{n}\frac{TP_i}{TP_i + FN_i} \tag{9}$$

and

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}. \tag{10}$$

TABLE 3: Impact of contextual information on the segmentation performance of three DeepLabV3+ backbones using the test set of AWE. Results are reported for models trained with (w Ctx.) and without (w/o Ctx.) context.

Model / Backbone	Setting	mIoU [%]	Acc [%]	Prec. [%]	Recall [%]	F1 [%]
DeepLabV3+ / ResNet [60]	w/o Ctx.	77.71	99.69	88.34	84.09	86.16
	w Ctx.	81.46	99.74	89.07	87.47	88.26
DeepLabV3+ / MobileNet [61]	w/o Ctx.	77.31	99.68	86.64	84.74	85.68
	w Ctx.	78.27	99.72	85.40	86.63	86.01
DeepLabV3+ / Xception [62]	w/o Ctx.	68.74	99.34	87.75	74.52	80.60
	w Ctx.	78.91	99.70	86.62	86.42	86.52

C. IMPLEMENTATION DETAILS

The experiments were conducted on a personal desktop computer with a GeForce Titan Xp with 12GiB of VRAM. For the training procedure, stochastic gradient descent (SGD) was used with a momentum of 0.9 and a weight decay of 5×10^{-4}. The batch size was set to 4 and the learning rate to 7×10^{-3} for all models. The training images were cropped to a fixed size of 512×512, and the average value computed over the whole training set was subtracted for each channel. The training was run for 50 epochs with the stopping criteria of loss value not decreasing anymore. The context provider of ContexedNet was implemented with $c_f = 19$ feature maps at the output. The code (written in PyTorch) used for the experiments is made publicly available to foster reproducibility from: http://awe.fri.uni-lj.si/.

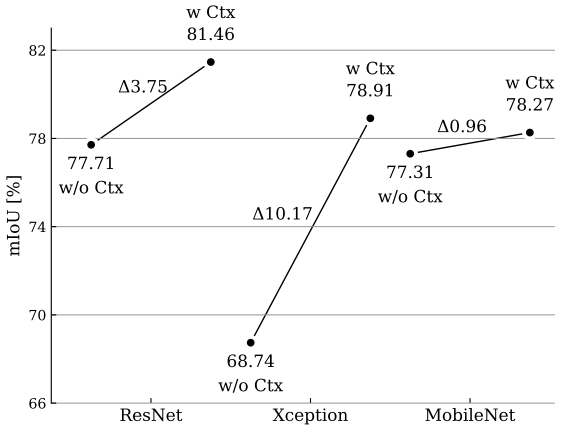


FIGURE 6: Impact of context on the three backbones considered in the experiments in terms of mIoU scores.

V. RESULTS

To demonstrate the merits of ContexedNet and capitalize on the importance of contextual information for the overall performance of the proposed ear detection solution, this section presents experimental results that: (i) highlight the impact of the proposed contextualization with three different baseline segmentation models, (ii) illustrate the effect of context on ear segmentation performance in a fine–grained analysis involving multiple covariates, (iii) present qualitative examples of successful and failed detections, (iv) analyze some of the framework's main characteristics, and (iv) compare the proposed approach to state–of–the–art solutions from the literature.

A. IMPACT OF CONTEXTUAL INFORMATION

The first series of experiments explores the impact of the context provider on the performance of ContexedNet's segmentation model. To this end, the DeepLabV3+ model [17] used in the segmentation path of ContexedNet is implemented using three different backbones, i.e., ResNet [60], MobileNet [61] and Xception [62]. Publicly available code is used as the basis for implementing these backbones[5].

1) Training Characteristics and Test Time Performance

In Figure 5 we visualize the training characteristics of the models trained with and without the context provider. As can be seen, all three backbone models exhibit significantly better convergence when used with contextual information. Given the same training data, the context–supported models not only converge faster, but (in most cases) also reach a better optimum than the models trained without context, as shown by the mIoU scores in Figure 5.

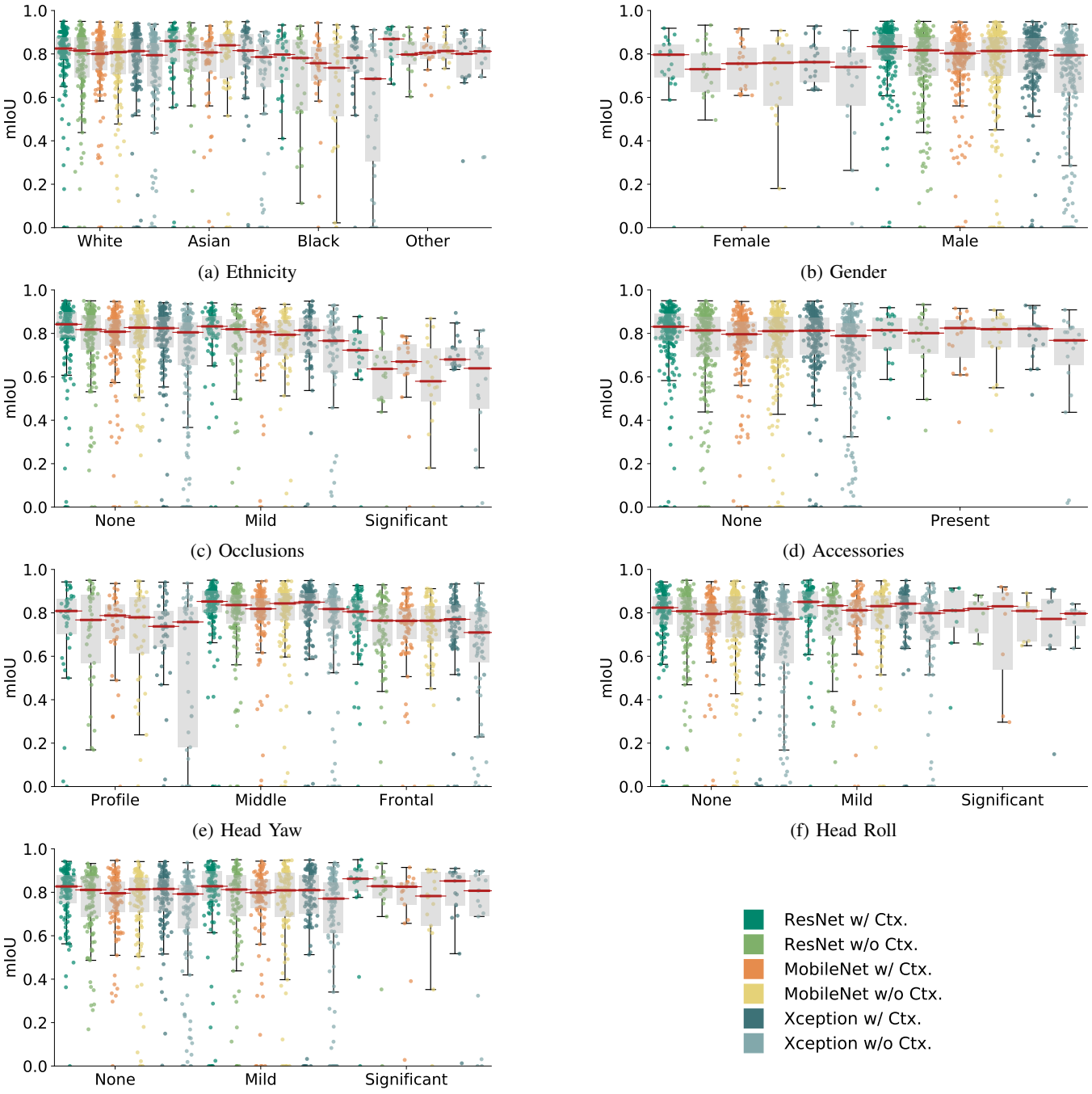
The overall processing time needed for training and testing of the models with our experimental hardware is given in Table 2. Note that training the context provider takes around a day. Once the model is trained and feature maps encoding face part locations are added as input to the segmentation model a gain of around 10 minutes is observed when training the context–aware segmentation models. At run–time, the additional processing needed to compute the contextual information results in an increase of the computational time of $2\times$ to $3\times$, and takes around $0.11s$ for the segmentation model with the ResNet backbone, $0.09s$ for the MobileNet backbone and $0.12s$ for the Xception backbone on average.

2) Performance Assessment

Next, we evaluate the three DeepLabV3+ backbone models on the test part of the AWE dataset with the goal of assessing the impact of contextual information on the overall segmentation performance. Again, backbones trained with and without contextual information are considered for this experiment.

The results in Table 3 show that context has a consid-

[5]Available from: https://github.com/jfzhang95/pytorch-deeplab-xception and https://github.com/switchablenorms/CelebAMask-HQ



(a) Ethnicity

(b) Gender

(c) Occlusions

(d) Accessories

(e) Head Yaw

(f) Head Roll

(g) Head Pitch

FIGURE 7: Fine–grained analysis of the impact of contextual information on the performance of three backbone models used for the implementation of DeepLabV3, i.e., ResNet, MobileNet and Xception. The models are trained with and without contextual information on the AWE dataset. Seven groups of covariates with the following number of images are considered: (a) Ethnicity (White - 458, Asian - 167, Black - 80, Other – 45), (b) Gender (Female – 69, Male – 681), (c) Occlussions (None – 488, Mild – 206, Significant – 56), (d) Accessories (None – 682, Present – 68), (e) Head yaw (Profile – 113, Middle – 429, Frontal – 208), (f) Head roll (None – 469, Mild – 246, Significant – 35), (g) Head pitch (None – 418, Mild – 298, Significant – 34). IoU scores are shown on the y–axes and covariate groups on the x–axes. Best viewed in color.

TABLE 4: Impact of contextual information on the segmentation bias across seven (demographic and non-demographic) covariates. Results are presented in terms of MAD scores, where smaller scores imply less biased results. Note that the integration of context reduces segmentation bias in the majority of cases, as also evidenced by the average MAD score.

Segmentation Backbone	Setting	Ethnicity	Gender	Occlusions	Accessories	Head Yaw	Head Roll	Head Pitch	Average
ResNet [60]	w Ctx.	0.0299	0.0260	0.0463	0.0010	0.0295	0.0162	0.0071	0.0223
	w/o Ctx.	0.0269	0.0252	0.0621	0.0047	0.0315	0.0158	0.0090	0.0250
MobileNet [61]	w Ctx.	0.0337	0.0200	0.0552	0.0099	0.0224	0.0237	0.0081	0.0247
	w/o Ctx.	0.0445	0.0387	0.0809	0.0067	0.0345	0.0130	0.0130	0.0330
Xception [62]	w Ctx.	0.0155	0.0158	0.0408	0.0137	0.0373	0.0274	0.0071	0.0225
	w/o Ctx.	0.0183	0.0161	0.0486	0.0055	0.0186	0.0351	0.0284	0.0244

erable impact on both mIoU as well as accuracy scores of all three tested models. The largest performance difference is observed with the Xception model, where the mIoU is improved by 10.17 percentage points through the contextualization, and the smallest with the MobileNet model with an improvement of 0.96 percentage points in terms of mIoU, as additionally illustrated in Figure 6. A jump of 3.75 percentage points is seen with the ResNet model, which also performs best overall among all tested backbones with an mIoU score of 81.46% when contextual information is included in the segmentation procedure. Consistent relative performance improvements are also observed for the tested backbone models when looking at the accuracy, precision, recall and F1 scores.

The presented results clearly show that contextual information is beneficial for ear segmentation and results in consistent performance improvements over context–free models. Additionally, performance gains are observed with all backbone models, suggesting that the proposed contextualization generalizes well over different CNN architectures.

3) Covariate Analysis

To further investigate the impact of contextual information, we conduct a fine–grained performance analysis on the test part of the AWE dataset. Specifically, we explore the segmentation performance of the three DeepLabV3+ backbone models, ResNet, Xception and MobileNet, trained with and without contextual information in the presence of different covariates. The results of this experiment are presented in the form of box–and–whiskers plots in Figure 7. Seven groups of covariates are considered, i.e., ethnicity, gender, presence of occlusions, presence of accessories, and head rotations in terms of yaw, roll and pitch.

Several interesting observations can be made from the presented results: (i) for an overwhelming majority of subgroups, the inclusion of contextual information consistently improves the median mIoU scores across all three backbones and (equally important) improves the distribution of the scores by reducing the dispersion over the test images, (ii) the contextualization has the biggest (positive) impact on the Xception backbone, followed in order by the ResNet and MobileNet models, where improvements are observed for the majority of subgroups considered, (iii) in absolute terms, the context–aware ResNet is again the most competitive among the tested backbones across all covariates, (iv) the integration of contextual information results in the biggest performance gains (on average) in the most challenging conditions, e.g., in the presence of significant occlusions (Figure 7c), as well as across different head rotations (Figures 7e to 7g), (v) performance gains are also observed across demographic factors, ethnicity and gender, where IoU scores are improved significantly for some of the subgroups that performed weaker without contextual information (Figures 7a and 7b).

4) Bias Analysis

The result, presented in the previous section, demonstrated the impact of contextual information on the performance of the segmentation model in terms of absolute gains. However, another critical issue with contemporary machine learning models is bias [68]–[72]. Machine learning models are expected to produce consistent results regardless of the demographic characteristics associated with the test images and to perform equally well for images with different non-demographic characteristics. To investigate the impact of the contextual information used in ContexedNet with respect to segmentation bias[6] mean absolute deviations (MAD) are computed across the covariate groups analyzed in Figure 7. Specifically, let $\mathcal{C}$ denote a given covariate class/group (e.g, ethnicity) and let mIoU_c represent the mIoU score associated with the c-th label from $\mathcal{C}$ (e.g., Asian) then the corresponding MAD can be defined as follows:

$$MAD = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}|mIoU_c - \overline{mIoU}|, \tag{11}$$

where $|\mathcal{C}|$ denotes the cardinality of $\mathcal{C}$, and $\overline{mIoU}$ stands for the mean mIoU score for the covariate class $\mathcal{C}$. Lower values of MAD indicate lower bias. MAD takes a value of 0 in the ideal case when no bias is present.

The MAD scores for the seven covariate groups analyzed are presented in Table 4. Note that the inclusion of contextual information significantly reduces the overall segmentation bias for the majority of image subgroups. The average MAD score for ResNet is reduced by 10.8%, by 25.2%

[6]When associated with demographic factors, bias is often related to the notion of fairness in machine learning. It is also often described with the term *differential outcome* to imply that different data characteristics may result in different performance [68].

FIGURE 8: Comparison of the segmentation path of ContexedNet trained with and without contextual information. Results are presented with the ResNet backbone model used for implementation of DeepLabV3. Predictions marked red correspond to results without contextual information, predictions in blue correspond to results produced with contextual information, and pink areas correspond to overlapping regions. The figure is best viewed in color.

for MobileNet and by 7.8%for Xception when context is used. This observation points to the fact that contextual information is not only useful to improve performance, but also contributes towards more consistent results across various image characteristics.

5) Qualitative Evaluation

The evaluations presented so far demonstrated the importance of contextual information for the overall segmentation performance of ContexedNet. Among the tested backbones, the ResNet model achieved the best overall performance and is, therefore, also used in most of the following experiments.

To further illustrate the value of contextual information, a comparison of the ResNet–based segmentation path trained with and without context is presented in Figure 8. This qualitative analysis is done with a few (challenging) test images collected from the web, so the test data is completely independent from the AWE dataset. Segmentation results produced by the model trained without context are shown in red, results with context in blue, and overlapping regions are shown in pink. As can be seen, the use of contextual information significantly improves performance. Without context, ear regions are often detected in semantically unreasonable areas that do resemble ears in terms of visual appearance, but are located in areas without meaningful context. With the integration of contextual cues such erroneous segmentations do not happen (or happen less often) due to the strong prior provided by the face part locations.

In Figure 9, a few additional example images are shown, where the context–free model completely fails to detect ear regions, while the proposed context–aware model not only successfully detects ear regions, but also generates high-quality segmentation masks that very well capture ear locations. We again attribute this behavior to the global approach used with ContexedNet, where semantically meaningful contextual information is exploited by the segmentation procedure instead of learning only from (spatially local) ear appearances.

B. CONTEXEDNET ANALYSIS

The second series of experiments analyzes some of the main characteristics of the proposed ContexedNet framework. Several experiments are presented, including: (i) an ablation study, (ii) an analysis of the impact of backbone models used for the implementation of ContexedNet, and (iii) an investigation into the use of face detection as a preprocessing step to ear segmentation.

1) Ablation Study

The proposed ContexedNet uses a two-path approach to segment ear regions from the input images. To demonstrate the importance of this two-path procedure, we conduct a simple ablation study and implement an additional single path model that predicts the ear region as well as all other face parts in a single computing step. This one-path model

FIGURE 9: Example images, where the context–free model completely fails, while the model trained with contextual information is still able to successfully detect ears and generate high–quality segmentation masks – shown in blue.

TABLE 5: Comparison of one-path and two-path approaches to context-aware ear segmentation on the test data of AWE. The one-path approach is implemented only with the Context Provider, the two-path approach is the proposed ContexedNet, which also offers superior performance.

Setting	mIoU [%]	Acc [%]	Prec. [%]	Recall [%]	F1 [%]
One-path	30.67	96.52	30.68	98.74	46.82
Two-path	81.46	99.74	89.07	87.47	88.26

essentially consists of only the context provider that in one of the output channels also produces segmentation maps of the ear region. Thus, the model still considers contextual information, but does not rely on a separate ear segmentation model when generating the final results. A comparison of the two-path approach of ContexedNet and the implemented one-path solution is presented in Table 5.

As can be seen, the complete two-path ContexedNet model convincingly outperforms the one-path procedure. While the simpler one-path approach has obvious run-time advantages due to the use of a single-step pipeline, it is only able to provide coarse segmentation results. Conversely, the proposed ContexedNet not only makes efficient use of the contextual information generated by the context provider, but also acts as a sort of *refinement network* for the output of the first path that produces finer and more accurate segmentations, as also illustrated in Figure 10.

FIGURE 10: Visual comparison of segmentation results produced by the one-path (i.e., the Context Provider - marked light blue), and the two-path (ContexedNet - marked magenta) models.

2) Backbone Evaluation

As suggested earlier, the contextualization proposed in this paper is general and can be used with any backbone model in either of the two paths of ContexedNet. We illustrate this flexibility by implementing the entire pipeline with a SegNet model and use SegNet for both, the context provider as well as the context-aware segmentation network. The SegNet based implementation of ContexedNet is compared to the best performing DeepLabV3+ based version (using ResNet) in Table 6. The results generated on the test part of AWE show that the proposed contextualization (marked w. Ctx.) contributes to considerable performance improvements regardless of the backbone model used. We observe a somewhat larger relative performance gain with SegNet, but in absolute terms the ContexedNet version implemented with DeepLabV3+ still yields the overall better results due to the superior baseline performance of the DeepLab model.

3) Context Exploration

ContexedNet uses contextual information in the form of face-part locations to improve segmentation performance. Additionally, the DeepLabV3+ based version also exploits atrous convolutions that capture spatial context to aid the segmentation procedure. However, existing ear detection

TABLE 6: Comparison of ContexedNet variants implemented with two different backbone models, i.e., DeepLabV3+ (ResNet based) and SegNet on the AWE test data. Note that regardless of the backbone model used, the proposed contextualization contributes to improved segmentation performance.

Backbone	Setting	mIoU [%]	Acc [%]	Prec. [%]	Recall [%]	F1 [%]
DeepLabV3+† [17]	w/o Ctx.	77.71	99.69	88.34	84.09	86.16
	w Ctx.	81.46	99.74	89.07	87.47	88.26
SegNet [73]	w/o Ctx.	37.81	98.60	46.55	59.66	52.30
	w Ctx.	48.61	98.91	57.89	71.64	64.04

†Results are reported for DeepLabV3+ with the ResNet backbone.

TABLE 7: Impact of face detection on segmentation performance. Results are reported on the AWE test data for models trained with (w Ctx.) and without (w/o Ctx.) contextual information in the form of face-part locations.

Backbone	Setting	mIoU [%]	Acc [%]	Prec. [%]	Recall [%]	F1 [%]
Complete Images	w/o Ctx.	77.71	99.69	88.34	84.09	86.16
	w Ctx.	81.46	99.74	89.07	87.47	88.26
Cropped Faces	w/o Ctx.	75.72	99.58	87.66	83.99	85.79
	w Ctx.	80.61	99.67	88.46	88.88	88.67

techniques typically rely on a separate face detection step to first constrain the spatial area in the input images before attempting ear detection/segmentation. This face detection step can be considered as another source of contextual information that restricts the spatial area of the input images that needs to be examined for the presence of ears. In the next experiment we, therefore, investigate whether face detection further contributes towards the performance of ContexedNet. To this end, we manually crop the face regions from the input images and train ContexedNet with cropped inputs. This procedure simulates face detection in an oracle type of setting, where perfect face detection results are assumed. We test the trained model with cropped test images from the AWE dataset and report results in Table 7. Here, results are again reported with and without the context provider for the DeepLabV3+ based version of ContexedNet.

Interestingly, restricting the search space of ContexedNet to the cropped facial area does not have a significant effect on performance. While minor differences in the individual performance scores are observed, these are very minute and have a limited impact on operational aspects of the segmentation model. When looking at the impact of the contextualization procedure, we see that the added information on face-part locations (marked w Ctx.) is beneficial even if the facial area is cropped. However, overall the added computational overhead and limited performance gains in general do not justify using a face detection approach as a preprocessing step to ear segmentation with ContexedNet. The proposed model alone is sufficient to ensure competitive performance, as shown by our experiments.

C. COMPARISON TO THE STATE–OF–THE–ART

In the last series of experiments, we compare ContexedNet to competing solutions from the literature on the AWE and UBEAR datasets. The ResNet-based DeepLabV3+ model is used as the backbone for ContexedNet's segmentation path due to its favorable performance compared to the two other backbones explored in the previous sections.

TABLE 8: Performance comparison with the state–of–the–art on the AWE dataset. All competing models are segmentation based. The results demonstrate the importance of of contextual information for segmentation–based ear detection.

Det./Seg. Approach	mIoU [%]	Acc [%]	Prec. [%]	Recall [%]	F1 [%]
SegNet [73]	37.81	98.60	46.55	59.66	52.30
PED–CED [1]	55.70	99.40	67.70	77.70	72.36
DeepLab [17]	77.71	99.68	88.34	84.09	86.16
ContexedNet (ours)	81.46	99.74	89.07	87.47	88.26

1) Results on the AWE Dataset

For the comparison on the AWE dataset, three state-of-the-art models are implemented, i.e., SegNet [73], PED–CED [1] and the DeepLab model from [17]. These models pose ear detection as a segmentation problem and are, therefore, directly comparable to the proposed ContexedNet – implemented with the ResNet-based DeepLabV3+ model for these experiments. The results in Table 8 show that all models result in comparable accuracy due to the impact of the majority class (i.e., the background) on this performance score. However, convincing improvements are observed when looking at the more informative mIoU scores and the precision, recall and F1 values, which are focused only on the ear segmentation performance and not the background. With these performance measures, ContexedNet significantly outperforms PED–CED and also ensures a considerable improvements over DeepLab, which represents a context–free segmentation model. These results clearly demonstrate the added value of contextual information for the task of ear detection/segmentation and the superiority of the proposed ContexedNet.

2) Results on the UBEAR Dataset

To further validate the performance of ContexedNet, we compare the model with competing solutions on the UBEAR dataset [18]. Specifically, we use the best performing segmentation–based approach from the experiments in Table 8, DeepLab, as well as two state–of–the–art bounding-box based ear detectors, i.e., MS–Faster R–CNN [7] and CED–Net [6]. Both MS–Faster R–CNN and CED-Net represent variants of the Faster R–CNN object detector. However, the latter also considers spatial context (as illustrated in Figure 1(a)) and is, therefore, context–aware, similarly to ContexedNet. Additionally, we also include results for the Single Shot Multi Box Detector (SSD) [74] and the original Faster R–CNN model [75], again trained for (bounding box) ear detection. Results for these two models are borrowed from [6]. MS–Faster R–CNN, CED–Net, SSD and Faster R–CNN return bounding boxes and not pixel–level

TABLE 9: Comparative evaluation on the UBEAR dataset. The results were generated in accordance with the experimental protocol presented in Table 1. Note that the reported results were computed based on bounding–boxes and not pixel-level segmentation masks (based on which ContexedNet was trained) to allow for a fair comparison.

Detection/Segmentation Approach	Accuracy [%]		Precision [%]		Recall [%]		F1 score [%]	
	ΔIoU= 0.6	ΔIoU= 0.7	ΔIoU= 0.6	ΔIoU= 0.7	ΔIoU= 0.6	ΔIoU= 0.7	ΔIoU= 0.6	ΔIoU= 0.7
MS–Faster R–CNN† [7]	‡98.22		‡99.55		‡98.66		‡99.10	
CED–Net† [6]	99.84	99.35	99.84	99.35	99.87	99.38	99.86	99.36
SSD† [6], [74]	92.17	85.88	96.78	90.18	98.32	91.62	97.55	90.89
Faster R–CNN† [6], [75]	92.09	82.74	96.51	86.71	96.12	86.35	96.31	86.53
DeepLab [17]	93.98	88.86	95.31	90.12	98.54	98.46	95.31	90.12
ContexedNet (ours)	95.51	91.20	95.94	91.61	99.53	99.51	95.94	91.61

† Results are borrowed from [7], [6]. ‡ IoU threshold not considered in [7]; objectness threshold of 0 was used. Color coding: Bound-box detection techniques, Segmentation-based techniques.

FIGURE 11: Example bounding-box ear detection results on sample images from the UBEAR dataset. Note that bounding boxes were fitted to the segmentation masks generated by ContexedNet. The blue annotations correspond to the ground truth, the red ones to the output of ContexedNet and the magenta annotations to the overlap between the two. The figure is best viewed in color.

segmentation masks, Table 9, therefore, reports detection-based performance scores computed based on bounding box information and not based on segmentation masks. Pixel–level accuracy, precision, recall and F1 scores are not reported, as they do not apply to this detection setting. To make the segmentation models, DeepLab and ContexedNet, comparable to the detection procedures, a bounding box is fitted to the generated segmentation masks prior to computing performance scores. Training and testing of the segmentation models is done in accordance with the experimental setup from Table 1, where half of the data is used for training and validation, and half for the final performance evaluation, similarly to [6]. Results are reported for two IoU thresholds, i.e., ΔIoU$= 0.6$ and ΔIoU$= 0.7$.

Table 9 shows that among the tested models CED–Net and MS–Faster R–CNN perform best in terms of the generated accuracy scores, which suggests that these models are highly successful in detecting ears in UBEAR images. The proposed ContexedNet also achieves highly competitive performance despite not being trained for bounding-box detection at all[7]. Our framework again benefits from the proposed contextualization and convincingly outperforms the context-free DeepLab model with respect to the accuracy score. We also observe superior performance when comparing ContexedNet to the SSD and Faster R-CNN (bounding-box) ear detectors, where our framework has a clear edge. Similar observations can also be made when looking at the precision, recall and F1 scores that again point to the impressive performance of ContexedNet. To put the reported quantitative results into perspective, we show in Figure 11 a few example detection results – with fitted bounding boxes for ContexedNet. Note how (despite the fitting procedure) the bounding-boxes correspond reasonably well to the annotated ground truth.

VI. CONCLUSION

In this paper, a novel context–aware ear detection framework, called ContexedNet, was presented. The framework exploits information on face-part locations to improve ear detection/segmentation performance and improves on existing segmentation–based solutions to ear detection by learning from contextual cues in addition to ear appearances. The model was tested in comprehensive experiments on the AWE and UBEAR datasets. Experimental results suggest that the use of contextual information not only improves detection performance compared to context–free models, but also that the contextualization has a beneficial effect on reducing segmentation bias across various (demographic and non–demographic) covariates. Additionally, the model was shown to ensure competitive performance when compared to state-of–the–art solutions from the literature both on AWE as well as UBEAR.

As part of our future work on this topic, we plan to strengthen the integration of the context provider in the overall processing pipeline (using multi–task learning, for example), so it is trainable in an end–to–end manner. Additionally, we plan to incorporate additional learning objectives and criteria that can further constrain the segmentation procedure. The developed detection approach will also be incorporated into an ear recognition system, where the

[7]Note again that bounding boxes were fitted to the generated segmentation masks. The correspondence with the ground truth annotations has, therefore, not been learned as with the competing detection techniques.

pixel–level output produced by ContexedNet will be used during feature learning.

ACKNOWLEDGEMENTS

This research was supported in parts by the ARRS Research Program P2-0250 (B) "Metrology and Biometric Systems" and the ARRS Research Program P2-0214 (A) "Computer Vision". We also thank NVIDIA for their donation of GPUs used for this work.

REFERENCES

[1] Ž. Emeršič, L. L. Gabriel, V. Štruc, and P. Peer, "Convolutional encoder–decoder networks for pixel-wise ear detection and segmentation," *IET Biometrics*, vol. 7, no. 3, pp. 175–184, 2018.

[2] Ž. Emeršič, V. Štruc, and P. Peer, "Ear Recognition: More Than a Survey," *Neurocomputing*, vol. 255, pp. 26–39, 2017.

[3] 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, 2012.

[4] Ž. Emeršič, B. Meden, P. Peer, and V. Štruc, "Evaluation and analysis of ear recognition models: Performance, complexity and resource requirements," *Neural Computing and Applications*, pp. 1–16, 2020.

[5] M. Bizjak, P. Peer, and Ž. Emeršič, "Mask r-cnn for ear detection," in *2019 42nd International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, May 2019, pp. 1624–1628.

[6] A. Kamboj, R. Rani, A. Nigam, and R. R. Jha, "Ced-net: context-aware ear detection network for unconstrained images," *Pattern Analysis and Applications*, pp. 1–22, 2020.

[7] Y. Zhang and Z. Mu, "Ear Detection under Uncontrolled Conditions with Multiple Scale Faster Region-Based Convolutional Neural Networks," *Symmetry*, vol. 9, no. 4, pp. 1–19, Apr. 2017.

[8] Ž. Emeršič, J. Križaj, V. Štruc, and P. Peer, "Deep Ear Recognition Pipeline," in *Recent Advances in Computer Vision*, M. Hassaballah and K. M. Hosny, Eds., 2019, vol. 804, pp. 333–362.

[9] R. Girshick, "Fast R-CNN," in *International Conference on Computer Vision*, 2015, pp. 1440–1448.

[10] S. Ren, K. He, R. B. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, pp. 1137–1149, 2017.

[11] K. He, G. Gkioxari, P. Dollár, and R. B. Girshick, "Mask R-CNN," *CoRR*, vol. abs/1703.06870, 2017.

[12] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, vol. 28, pp. 91–99.

[13] A. Bochkovskiy, C.-Y. Wang, and H.-Y. M. Liao, "Yolov4: Optimal speed and accuracy of object detection," *arXiv preprint arXiv:2004.10934*, 2020.

[14] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," *arXiv preprint arXiv:2103.14030*, 2021.

[15] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[16] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Conference on Computer Vision and Pattern Recognition*, no. 2, 2017, p. 5.

[17] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *European Conference on Computer Vision*, 2018.

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

[19] A. Kamboj, R. Rani, and A. Nigam, "A comprehensive survey and deep learning-based approach for human recognition using ear biometric," *The Visual Computer*, pp. 1–34, 2021.

[20] B. Arbab-Zavar and M. S. Nixon, "On Shape-Mediated Enrolment in Ear Biometrics," in *International Symposium on Visual Computing*, 2007, pp. 549–558.

[21] P. Chidananda, P. Srinivas, K. Manikantan, and S. Ramachandran, "Entropy-cum-Hough-transform-based ear detection using ellipsoid particle swarm optimization," *Machine Vision and Applications*, vol. 26, no. 2, pp. 185–203, 2015.

[22] S. Attarchi, K. Faez, and A. Rafiei, "A New Segmentation Approach for Ear Recognition," in *International Conference on Advanced Concepts for Intelligent Vision Systems*, 2008, pp. 1030–1037.

[23] S. Ansari and P. Gupta, "Localization of Ear Using Outer Helix Curve of the Ear," in *International Conference on Computing: Theory and Applications*, 2007, pp. 688–692.

[24] S. Prakash, U. Jayaraman, and P. Gupta, "Ear Localization from Side Face Images using Distance Transform and Template Matching," in *Workshops on Image Processing Theory, Tools and Applications*, 2008, pp. 1–8.

[25] ——, "Connected Component based Technique for Automatic Ear Detection," in *International Conference on Image Processing*, 2009, pp. 2741–2744.

[26] S. Prakash and P. Gupta, "An efficient ear localization technique," *Image and Vision Computing*, vol. 30, no. 1, pp. 38–50, 2012.

[27] S. Prakash, U. Jayaraman, and P. Gupta, "A Skin-Color and Template Based Technique for Automatic Ear Detection," in *International Conference on Advances in Pattern Recognition*, 2009, pp. 213–216.

[28] A. H. Cummings, M. S. Nixon, and J. N. Carter, "A Novel Ray Analogy for Enrolment of Ear Biometrics," in *International Conference on Biometrics: Theory Applications and Systems*, 2010, pp. 1–6.

[29] N. K. A. Wahab, E. E. Hemayed, and M. B. Fayek, "HEARD: An automatic human EAR detection technique," in *International Conference on Engineering and Technology*, 2012, pp. 1–7.

[30] A. Pflug, A. Winterstein, and C. Busch, "Robust Localization of Ears by Feature Level Fusion and Context Information," in *International Conference on Biometrics*, 2013, pp. 1–8.

[31] Z. Zou, Z. Shi, Y. Guo, and J. Ye, "Object detection in 20 years: A survey," *arXiv preprint arXiv:1905.05055*, 2019.

[32] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International Journal of Computer Vision*, vol. 128, no. 2, pp. 261–318, 2020.

[33] S. M. Islam, M. Bennamoun, and R. Davies, "Fast and Fully Automatic Ear Detection Using Cascaded Adaboost," in *Workshop on Applications of Computer Vision*, 2008, pp. 1–6.

[34] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features," in *Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001, pp. I–511–I–518.

[35] A. Abaza, C. Hebert, and M. A. F. Harrison, "Fast Learning Ear Detection for Real-time Surveillance," in *International Conference on Biometrics: Theory Applications and Systems*, 2010, pp. 1–6.

[36] H. Liu and D. Liu, "Improving adaboost ear detection with skin-color model and multitemplate matching," in *International Conference on Computer Science and Information Technology*, 2010.

[37] L. Yuan and F. Zhang, "Ear detection based on improved adaboost algorithm," in *International Conference on Machine Learning and Cybernetics*, vol. 4, 2009, pp. 2414–2417.

[38] University of Notre Dame, "Face Database," https://cvrl.nd.edu/projects/data/, 2015 (estimated), visited on 2018-12-25.

[39] S. El-Naggar, A. Abaza, and T. Bourlai, "Ear Detection in the Wild Using Faster R-CNN Deep Learning," in *International Conference on Advances in Social Networks Analysis and Mining*, Aug. 2018, pp. 1124–1130.

[40] A. Tomczyk and P. S. Szczepaniak, "Ear detection using convolutional neural network on graphs with filter rotation," *Sensors*, vol. 19, no. 24, p. 5510, 2019.

[41] 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, 2019.

[42] X. Tang, D. K. Du, Z. He, and J. Liu, "Pyramidbox: A context-assisted single shot face detector," in *European Conference on Computer Vision*, 2018, pp. 797–813.

[43] M. Najibi, P. Samangouei, R. Chellappa, and L. S. Davis, "Ssh: Single stage headless face detector," in *International Conference on Computer Vision*, 2017, pp. 4875–4884.

[44] O. Marques, E. Barenholtz, and V. Charvillat, "Context modeling in computer vision: techniques, implications, and applications," *Multimedia Tools and Applications*, vol. 51, no. 1, pp. 303–339, 2011.

[45] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie, "Objects in context," in *International Conference on Computer Vision*, 2007, pp. 1–8.
[46] C. Galleguillos and S. Belongie, "Context based object categorization: A critical survey," *Computer Vision and Image Understanding*, vol. 114, no. 6, pp. 712–722, 2010.
[47] J. Fan, H. Song, K. Zhang, Q. Liu, F. Yan, and W. Lian, "Real-time manifold regularized context-aware correlation tracking," *Frontiers of Computer Science*, vol. 14, no. 2, pp. 334–348, 2020.
[48] M. Mueller, N. Smith, and B. Ghanem, "Context-aware correlation filter tracking," in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1396–1404.
[49] A. Ross, S. Banerjee, C. Chen, A. Chowdhury, V. Mirjalili, R. Sharma, T. Swearingen, and S. Yadav, "Some research problems in biometrics: The future beckons," in *International Conference on Biometrics*, 2019, pp. 1–8.
[50] M. Nappi, S. Ricciardi, and M. Tistarelli, "Context awareness in biometric systems and methods: State of the art and future scenarios," *Image and Vision Computing*, vol. 76, pp. 27–37, 2018.
[51] K. Grm, W. J. Scheirer, and V. Štruc, "Face hallucination using cascaded super-resolution and identity priors," *IEEE Transactions on Image Processing*, vol. 29, pp. 2150–2165, 2019.
[52] B. Zhuang, L. Liu, C. Shen, and I. Reid, "Towards context-aware interaction recognition for visual relationship detection," in *International Conference on Computer Vision*, 2017, pp. 589–598.
[53] Y. Chen, X. Zhu, and S. Gong, "Instance-guided context rendering for cross-domain person re-identification," in *International Conference on Computer Vision*, 2019, pp. 232–242.
[54] R. Kosti, J. M. Alvarez, A. Recasens, and A. Lapedriza, "Emotion recognition in context," in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1667–1675.
[55] P. Rot, Ž. Emeršič, V. Struc, and P. Peer, "Deep multi-class eye segmentation for ocular biometrics," in *IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, 2018, pp. 1–8.
[56] J. Lozej, B. Meden, V. Struc, and P. Peer, "End-to-end iris segmentation using u-net," in *IEEE International Work Conference on Bioinspired Intelligence (IWOBI)*, 2018, pp. 1–6.
[57] P. Rot, M. Vitek, K. Grm, Ž. Emeršič, P. Peer, and V. Štruc, "Deep sclera segmentation and recognition," in *Handbook of vascular biometrics*. Springer, Cham, 2020, pp. 395–432.
[58] C.-H. Lee, Z. Liu, L. Wu, and P. Luo, "Maskgan: Towards diverse and interactive facial image manipulation," in *Conference on Computer Vision and Pattern Recognition*, 2020.
[59] J. Lozej, D. Štepec, V. Štruc, and P. Peer, "Influence of segmentation on deep iris recognition performance," in *International Workshop on Biometrics and Forensics (IWBF)*, 2019, pp. 1–6.
[60] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*, 2016, pp. 630–645.
[61] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, "MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications," *CoRR*, vol. abs/1704.04861, 2017.
[62] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1251–1258.
[63] C. Wang, Y. Wang, K. Zhang, J. Muhammad, T. Lu, Q. Zhang, Q. Tian, Z. He, Z. Sun, Y. Zhang *et al.*, "Nir iris challenge evaluation in non-cooperative environments: Segmentation and localization," in *IEEE International Joint Conference on Biometrics (IJCB)*, 2021, pp. 1–10.
[64] J. Šircelj, T. Oblak, K. Grm, U. Petković, A. Jaklič, P. Peer, V. Štruc, and F. Solina, "Segmentation and recovery of superquadric models using convolutional neural networks," in *Computer Vision Winter Workshop*, 2020.
[65] A. Das, U. Pal, M. A. Ferrer, M. Blumenstein, D. Štepec, P. Rot, Ž. Emeršič, P. Peer, V. Štruc, S. A. Kumar *et al.*, "Sserbc 2017: Sclera segmentation and eye recognition benchmarking competition," in *IEEE International Joint Conference on Biometrics (IJCB)*, 2017, pp. 742–747.
[66] M. Vitek, A. Das, Y. Pourcenoux, A. Missler, C. Paumier, S. Das, I. D. Ghosh, D. R. Lucio, L. A. Z. Jr., D. Menotti, F. Boutros, N. Damer, J. H. Grebe, A. Kuijper, J. Hu, Y. He, C. Wang, H. Liu, Y. Wang, Z. Sun, D. Osorio-Roig, C. Rathgeb, C. Busch, J. Tapia, A. Valenzuela, G. Zampoukis, L. Tsochatzidis, I. Pratikakis, S. Nathan, R. Suganya, V. Mehta, A. Dhall, K. Raja, G. Gupta, J. N. Khiarak, M. Akbari-Shahper, F. Jaryani, M. Asgari-Chenaghlu, R. Vyas, S. Dakshit, S. Dakshit, P. Peer, U. Pal, and V. Štruc, "Ssbc 2020: Sclera segmentation benchmarking competition in the mobile environment," in *International Joint Conference on Biometrics (IJCB 2020)*, 2020, pp. 1–10.
[67] A. Das, U. Pal, M. A. Ferrer, M. Blumenstein, D. Štepec, P. Rot, Žiga Emeršič, P. Peer, and V. Štruc, "Ssbc 2018: Sclera segmentation benchmarking competition," in *International Conference on Biometrics (ICB)*, 2018.
[68] J. J. Howard, Y. B. Sirotin, and A. R. Vemury, "The effect of broad and specific demographic homogeneity on the imposter distributions and false match rates in face recognition algorithm performance," in *International Conference on Biometrics Theory, Applications and Systems*, 2019, pp. 1–8.
[69] A. Puc, V. Štruc, and K. Grm, "Analysis of race and gender bias in deep age estimation models," in *European Signal Processing Conference*, 2021, pp. 830–834.
[70] K. Krishnapriya, V. Albiero, K. Vangara, M. C. King, and K. W. Bowyer, "Issues related to face recognition accuracy varying based on race and skin tone," *IEEE Transactions on Technology and Society*, vol. 1, no. 1, pp. 8–20, 2020.
[71] V. Albiero, K. KS, K. Vangara, K. Zhang, M. C. King, and K. W. Bowyer, "Analysis of gender inequality in face recognition accuracy," in *Winter Conference on Applications of Computer Vision Workshops*, 2020, pp. 81–89.
[72] P. Drozdowski, C. Rathgeb, A. Dantcheva, N. Damer, and C. Busch, "Demographic bias in biometrics: A survey on an emerging challenge," *IEEE Transactions on Technology and Society*, vol. 1, no. 2, pp. 89–103, 2020.
[73] V. Badrinarayanan, A. Handa, and R. Cipolla, "SegNet: A Deep Convolutional Encoder-Decoder Architecture for Robust Semantic Pixel-Wise Labelling," *CoRR*, vol. abs/1505.07293, 2015.
[74] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*, 2016, pp. 21–37.
[75] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2016.

ŽIGA EMERŠIČ a teaching assistant and a PhD candidate at the Faculty of Computer and Information Science, University of Ljubljana, Slovenia. Žiga's research primarily focuses on deep learning and ear biometrics. He co-authored more than 40 research papers, received multiple awards for teaching and research. He co-organized the first ear recognition competitions: the Unconstrained Ear Recognition Challenge 2017 and 2019 and co-organized one of the first machine–learning summer schools in Central America.

DIEGO SUŠANJ is a teaching assistant and a PhD candidate of Computer science at the Faculty of Engineering, University of Rijeka, Croatia. He received his B.S. degree in computer engineering in 2013. and his M.S. degree in computer engineering in 2015 from the University of Rijeka, Faculty of Engineering. His research interests are in the fields of computer vision, machine learning, image processing, and embedded systems. He is one of the founders of the Riteh Drone Team and a member of the Royal Institute of Navigation.

BLAŽ MEDEN is a PhD candidate and a teaching assistant at the Faculty of Computer and Information Science, University of Ljubljana, Slovenia. He received his BSc and MSc from the Faculty of Computer and Information Science and is currently working on his PhD in computer science at the University of Ljubljana. His PhD research is focused on facial privacy protection and generative deep learning approaches. More broadly, he is also interested in image based biometrics, image processing, and computer vision. Blaž also reviewed for a number of conferences and journals, such as IEEE Access, IET Biometrics, Entropy, IMAVIS, CVWW, ISPA, CSAE and IWOBI.

PETER PEER is a Professor of computer science at the University of Ljubljana, Slovenia, where he heads the Computer Vision Laboratory, coordinates the double degree study program with the Kyungpook National University, South Korea, and serves as a vice-dean for economic affairs. He received his doctoral degree in computer science from the University of Ljubljana in 2003. Within his post-doctorate he was an invited researcher at CEIT, Donostia – San Sebastian, Spain. His research interests include biometrics, color constancy, image segmentation, detection, recognition and real-time computer vision applications. He participated in several national and EU funded R&D projects and published more than 100 research papers in leading international peer reviewed journals and conferences. He is co-organizer of the Unconstrained Ear Recognition Challenge and Sclera Segmentation Benchmarking Competition. He serves as an Associated Editor of IEEE Access and IET Biometrics. He is a member of the EAB, IAPR and IEEE, where he also served as a chairman of the Slovenian IEEE Computer chapter for four years.

VITOMIR ŠTRUC is an Associate Professor at the University of Ljubljana, Slovenia. He received his doctoral degree from the Faculty of Electrical Engineering in Ljubljana in 2010. Vitomir's research interests include problems related to biometrics, computer vision, image processing, pattern recognition and machine learning. He (co-)authored more than 100 research papers for leading international peer reviewed journals and conferences in these and related areas. He served in different capacities on the organizing committees of several top-tier vision conferences, including IEEE Face and Gesture, ICB, WACV and others and is currently a program co-chair for the 2020 International Joint Conference on Biometrics (IJCB). Vitomir is a Senior Area Editor for the IEEE Transactions on Information Forensics and Security, and an Associate Editor for Pattern Recognition, Signal Processing, and IET Biometrics. He served as an Area Chair for WACV 2018, 2019, 2020, ICPR 2018, Eusipco 2019 and FG 2020. Dr. Struc is a member of the IEEE, IAPR, EURASIP, Slovenia's national contact point for the EAB and the current president of the Slovenian Pattern Recognition Society (Slovenian branch of IAPR).