PupilTAN: A Few-Shot Adversarial Pupil Localizer

Nikolaos Poulopoulos, Emmanouil Z. Psarakis and Dimitrios Kosmopoulos
Computer Engineering and Informatics, University of Patras, Greece
(npoul,psarakis)@ceid.upatras.gr, dkosmo@upatras.gr

Abstract

The eye center localization is a challenging problem faced by many computer vision applications. The challenges typically stem from the scene variability, such as, the wide range of shapes, the lighting conditions, the view angles and the occlusions. Nowadays, the increasing interest on deep neural networks requires a large volume of training data. However, a significant issue is the dependency on labeled data, which are expensive to obtain and susceptible to errors. To address these issues, we propose a deep network, dubbed PupilTAN, that performs image-to-heatmap Translation and an Adversarial training framework that solves the eye localization problem in a few-shot unsupervised way. The key idea is to estimate, by using only a few ground-truth shots, the heatmaps centers' pdf and use it as a generator to create random heatmaps that follow the same probability distribution of the real ones. We showcase that training the deep network with these artificial heatmaps in an adversarial framework not only makes us less dependent on labeled data, but also leads to a significant accuracy improvement. The proposed network achieves real-time performance in a general-purpose computer environment and improves the state-of-the-art accuracy for both MUCT and BioID datasets, even compared with supervised techniques. Furthermore, our model is robust even in the case of reducing its size of up to 1/16 of the original network (0.2M parameters), demonstrating comparable accuracy to the state-of-the-art with high practical value to real-time applications.

1. Introduction

The tremendous progress of computer systems over the last decades and their penetration in almost every aspect of the human life has inevitably induced a growing interest in improving the human-computer interaction (HCI). Systems that exploit the eye gaze, offer a convenient and natural mean of interaction without the requirement of physical contact. Eyes constitute the most distinctive features of the human face, while the iris positions with respect to the head pose and gaze are significant sources of information regarding the cognitive and affective state of human beings. Specifically, the information about the location of the eye centers is commonly used in applications such as face alignment, control devices for disabled people and user attention (e.g., driving and marketing) [1, 2]. Moreover, the eye center coordinates can be exploited to estimate the gaze by transforming the gaze angles (roll, yaw) to a 3D gaze vector [3]. Despite the active research in this field, precise eye center localization and tracking remains a challenging problem due to many limitations that downgrade the accuracy of the detected eye centers. These limitations are related to the great variety in shape and colour of human eyes, the eye state (open or closed), the iris direction, the facial expressions, the head pose etc. The localization accuracy can be also reduced under the presence of occlusions from hair, glasses, reflections and shadows and is strongly affected by the lighting conditions and the camera resolution.

Precise eye localization and tracking becomes even more challenging in real-time applications, where the need of real-time performance is crucial. Obtaining high-quality data to train supervised eye localizers constitutes an expensive and challenging task. Moreover, unintentionally and often unavoidably, labels are subject to human error (i.e., inaccurate ground-truth labels). On the other hand, there is a plenty of unlabeled eye data available for free.

In this paper, we introduce a novel framework, PupilTAN, that tries to solve the eye localization problem in an unsupervised way. Unsupervised learning is a type of algorithm that exhibits self-organization to capture the hidden patterns contained in unlabeled data. In contrast to supervised learning, which intends to infer a conditional on the label of the input data, unsupervised learning intends to infer an a-priori probability distribution.

In this context, we consider the eye localization problem as an image-to-heatmap regression and exploit the special form of the desired heatmaps. Specifically, we treat it as a 2-D isotropic gaussian kernel with constant standard deviation; its center is considered to be a normal random variable, whose *pdf* is estimated by using a small sample of the available ground-truth. The adversarial framework that we pro-

pose in the sequel exploits that knowledge and aims to train in an unsupervised way a translator network to capture the probability density of the incoming data. To the best of our knowledge, this is one of the first attempts, if not the first, in the literature that an adversarial framework is adopted to solve the eye localization problem, in an unsupervised manner. The main contributions of this work are summarized as follows:

- A novel adversarial framework for unsupervised eye localization.
- Superior accuracy over the state-of-the-art techniques in two publicly available databases.
- Significant reduction of network size with high practical value in real-time applications.

2. Related Work

In this section, we review relevant works on eye center localization and generative adversarial networks (GANs). **Eye localization methods**, can be roughly divided into the following two main categories:

- Feature-based methods and
- Appearance-based methods.

Feature-based methods use a priori knowledge to detect candidate eye centers from simple pertinent features based on shape, geometry, color and symmetry. These features are obtained from the application of specific filters on the image and don't require any learning or model-fitting techniques. The idea of isophote curvatures was proposed by Valenti et al. [4] as a voting scheme for detecting eye locations. Radial symmetry operators have also been employed for eye detection; they are typically combined with other operators [5, 6]. In works [7, 8] a Modified Fast Radial Symmetry Transform (MFRST) was proposed. It emphasized on the shape of the iris and combined the edge information that results from an edge-preserving filtering and the intensity information, in order to find shapes with high radial symmetry. In general, appearance-based methods employ a prior model of the eye holistic appearance and surrounding structures and try to detect the location of the eyes by fitting the trained model. For this purpose, many machine learning algorithms have been proposed such as Bayesian [9] and hidden Markov models (HMMs) [10], support vector machines (SVM) [11] and AdaBoost [12]. Markus et al. [13] localized the eye pupil by using an ensemble of randomized regression trees. Convolutional Neural Networks (CNNs) have recently attracted interest in their use as eye detectors. In Fuhl's [14] research, coarse to fine pupil localization was carried out using two similar convolutional neural networks. The first one provided a coarse position of the pupil while the second refined that position using smaller subregions as input. Li et al. in [15] also proposed a two-stage CNN to determine the most likely eye regions and to locate their centers. Deep CNNs have also achieved several improvements over the last years. Xia et al. [16] proposed a FCN with a large kernel convolutional block to localize the eye centers using heatmaps. In [17] and [18] a deep FCN pipeline was proposed using heterogenous CNN models trained to detect the face, remove the eye glasses, extract the facial landmarks and finally localize the pupil centers.

Generative Adversarial Networks, recently introduced in [19], aim to discover the underlying distribution from large amounts of data. Such models have been used to several tasks, like image generation [20], image composition [21], text-to-photo translation [22] and image-to-image translation [23].

The task of image-to-image translation involves learning of how to map a given source image to a specific target image. Learning the mapping from one visual representation to another requires an understanding of underlying features that are shared between these representations [24]. These approaches can be further divided into supervised and unsupervised ones. A supervised method requires a set of paired images in different domains and the model learns the probability distribution from one domain to another. Pix2Pix [25] was a supervised image-to-image translation approach based on a conditional generative adversarial network. The Generator used a "U-Net" like architecture and the Discriminator a convolutional based "PatchGAN" as classifier. Unsupervised image-to-image translation aims at learning the mapping between two or more domains without paired images and it has recently been explored intensively due its ability to learn the cross-mapping in the image-to-image translation problem. CycleGAN [23] aimed to learn the mapping between a set of unpaired images from two different domains. Its architecture was based on a symmetric structure of two Translators and two Discriminators and performed two mappings: the forward cycle mapping from the input domain to the target domain and the backward one. Robinson et. al [26] proposed a laplacian facial landmark localizer based on image-to-heatmap translation and improved its model accuracy using an adversarial framework trained on unlabeled data. Despite the growing interest, unsupervised localization remains challenging due to the difficulty to localize objects without annotations.

Our work tries to overcome this obstacle by transforming the eye localization problem to an image-to-heatmap translation and by training a generative adversarial network with random artificial heatmaps derived from the same pdf as the real ones. We believe that in the case of image-to-heatmap translation, the pdf of the paired heatmaps can be estimated, initially by using only a few ground-truth samples, and then training will be performed by using unlabeled images in an

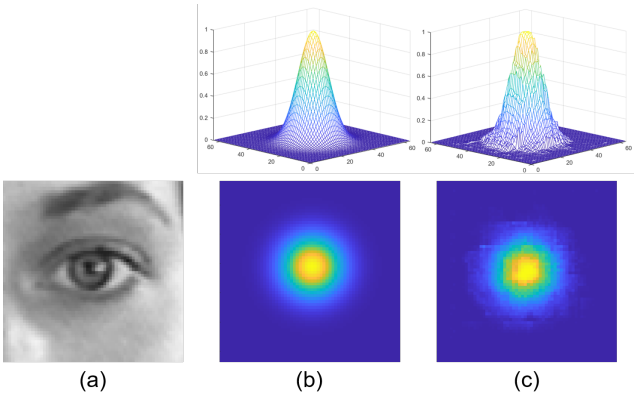

Figure 1. The eye's region around its center (a), its idealization modelled by a heatmap with center at the pixel where the real eye's center is located and std equal to the radius of the circle that covers the iris area (b), and the fake heatmap achieved by PupilTAN (c)

unsupervised fashion.

3. The Proposed Solution

In this section, we are going to give a detailed description of the proposed solution to the eye localization problem. To this end, we begin with the reformulation of the problem at hand into an image-to-heatmap regression one. In order to achieve our goal we considered an idealization of the region around the human eye center that can be modeled by a heatmap whose center coincides with the eye center and its standard deviation is controlled by the size of the iris. Our proposition is verified by the content of Figure 1.(a-c) where we can see an example.

In particular, we reformulate the eye localization problem into an image-to-heatmap regression problem and exploit the special form of the desired heatmaps. We demonstrate that instead of using the paired ground-truth to train an image-to-heatmap localizer, we can use randomly created heatmaps from the same pdf. Specifically, we treat them as a 2-D Gaussian kernel with constant standard deviation and its center as a normal random variable, whose *pdf* is estimated by using a small number of the available ground-truth data.

Based on our proposition, an adversarial framework is proposed that exploits the apriori knowledge and aims to train a deep neural network in an unsupervised way, using only a few ground-truth samples.

Finally, we are going to present the implementation details including the network architecture and the training procedure.

3.1. Preprocessing

To transform the eye center localization problem to an image-to-heatmap regression one, we propose the following preprocessing steps. Specifically, in every image of the training set, in the first stage of the pipeline the face is detected using the real-time face detector proposed by Viola and Jones [27] while in the second one, the two eye Regions of Interest (ROIs) are selected, based on the face geometry [7],[8]. Then, each ROI is resized to 64×64 pixels and transformed to a grayscale image in order to feed the translator. For every such image, the translator aims to predict a heatmap, of the same size, with its values indicating the per-pixel confidence of the location of the eye center. The position where its maximum value is attained corresponds to the predicted eye center coordinates.

It is clear that since for each image there is a ground-truth eye center $\mathbf{x}_{gt}$, we can derive the corresponding heatmap by using a kernel function, like the Gaussian one, as follows:

$$H(\mathbf{x}|\mathbf{x}_{gt}) = e^{\frac{||(\mathbf{x}-\mathbf{x}_{gt}||_2^2}{2\sigma^2}} \tag{1}$$

where, $\mathbf{x}$ belongs in the ROI and σ is the standard deviation of the kernel that determines the width of the heatmap. Due to the above mentioned preprocessing, the size of the iris has small variation and its approximate size can be inferred from the size of the detected face [6]. We set this hyperparameter to $\sigma = 7$, which represents the expected iris radius.

Note that given a large set of annotated images by following such an approach we create a proper supervised framework for the training of a deep neural network to solve the eye center localization problem. However, this in turn demands the existence of a large set of annotated images (i.e., for each image of the set a ground-truth eye center $\mathbf{x}_{gt}$ must be given) a fact that constitutes an obstacle for supervised training. To overcome this obstacle and transform the supervised training framework to a few-shot unsupervised one, we are going to treat the $\mathbf{x}_{gt}$ as a normally distributed Random Variable (RV) $\mathcal{X}_c$ whose the parameters $\mathbf{m}_c, \Sigma_c$ we are going to estimate by using a small sample of the given ground-truth eye centers. Our claim is based on the observation that the eye centers are normally distributed (Figure 2). This is exactly our goal in the next paragraph.

Estimating the Parameters $\mathbf{m}_c, \Sigma_c$

Thus, our objective is to estimate the parameters of a 2-D Gaussian function. To this end, let us consider that the following set of realizations of the above mentioned RV is given:

$$\mathcal{S}_c = \{\mathbf{x}_n\}_{n=1}^{N}. \tag{2}$$

By using this small sample of the ground-truth eye centers, we can estimate the parameters of the desired pdf by using

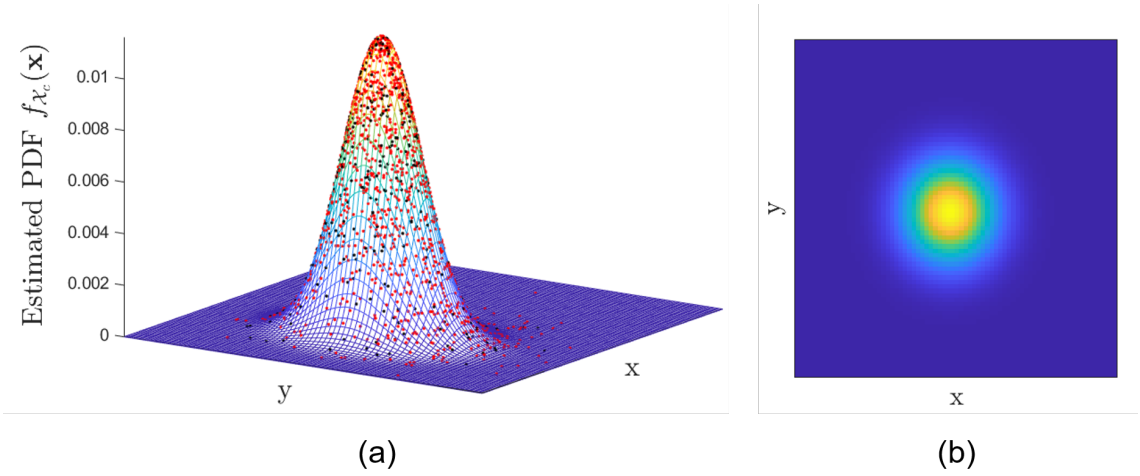

Figure 2. The Maximum Likelihood Estimated pdf $f_{\mathcal{X}_c}(\mathbf{x}_c)$ of the Random Variable $\mathcal{X}_c$ resulting from a small sample of 128 black dotted ground-truth eye centers of the face database BioID and the remaining red dotted ground-truth centers superimposed onto the estimated pdf (a). Top view of the $f_{\mathcal{X}_c}(\mathbf{x}_c)$ uncovering its isotropic nature (b)

the following maximum likelihood estimators:

$$\hat{\mathbf{m}}_c = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n \tag{3}$$

$$\hat{\Sigma}_c = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\hat{\mathbf{m}}_c)(\mathbf{x}_n-\hat{\mathbf{m}}_c)^T. \tag{4}$$

Having estimated the aforementioned parameters and with $|A|$ denoting the determinant of matrix$A$, we can use the following pdf:

$$f_{\mathcal{X}_c}(\mathbf{x}_c) = \frac{1}{2\pi|\hat{\Sigma}_c|^{\frac{1}{2}}}e^{-\frac{1}{2}(\mathbf{x}_c-\hat{\mathbf{m}}_c)^T\,\hat{\Sigma}_c^{-1}(\mathbf{x}_c-\hat{\mathbf{m}}_c)} \tag{5}$$

as a generator of realizations $\mathbf{x}_c$ of the RV $\mathcal{X}_c$ to define, using the relation:

$$H(\mathbf{x}|\mathbf{x}_c) = e^{\frac{||\mathbf{x}-\mathbf{x}_c||_2^2}{2\sigma^2}} \tag{6}$$

heatmaps that can be used for the unsupervised training of a network.

In Figure 2.(a) the Maximum Likelihood Estimated pdf $f_{\mathcal{X}_c}(\mathbf{x}_c)$ of the Random Variable $\mathcal{X}_c$ obtained from a small sample of 128 ground-truth eye centers of the face database BioID, shown as black dots in the figure, is depicted. Note that the remaining, shown with red dots, ground-truth centers are perfectly fitted to the estimated pdf. The top view of the resulting pdf shown in Figure 2.(b) uncovers its isotropic nature, a fact that simplifies the estimation problem and restricts the number of the ground-truth needed for its solution.

3.2. Unsupervised Eye Localization

Let us consider the following set of training images:

$$\mathcal{S}_{\mathcal{I}} = \{I_k\}_{k=1}^{K} \tag{7}$$

with each member I_k of this set constituting a realization of the Random Variable $\mathcal{I}$, i.e.:

$$I_k \sim f_{\mathcal{I}}(I) \tag{8}$$

where *pdf* $f_{\mathcal{I}}(I)$ is unknown. Having created the mechanism to generate heatmaps, we can form a large set of samples, let us denote it by:

$$\mathcal{S}_{\mathcal{H}} = \{H(\mathbf{x}_n|\mathbf{x}_c)\}_{n=1}^{M},\ M >> N, \tag{9}$$

that can be used for the training of the Translational Adversarial Network (TAN) shown in Figure 3. The entire network is composed by the Translator and the Discriminator subnetworks. In such a deep architecture during the training phase, the goal of the translational part of the network $H(\mathbf{x}|I(\mathbf{x}))) = T(I(\mathbf{x});\theta)$, is to model through a set of parameters θ the unknown *pdf* $f_{\mathcal{H}|\mathcal{I}}(H)$ of the heatmaps $H(\mathbf{x}|I(\mathbf{x}))$.

The translator tries to confuse the discriminator by generating images as plausible as possible, while, at the same time, the discriminator through a set of parameters ϑ in a fully adversarial way tries to distinguish the fake translated heatmaps from the real ones, thus enforcing the translator to produce heatmaps that are as close they can be to the "real" ones $H \sim f_{\mathcal{H}}(H)$. To this end, we are going to use the following adversarial loss function proposed by [19]:

$$\begin{aligned}\mathcal{L}(\theta,\vartheta) = {} & \mathbb{E}_{\mathcal{H}\sim f_{\mathcal{H}}(H)}[log(D(H(\mathbf{x}|\mathbf{x}_c)),\vartheta))] + \\ & \mathbb{E}_{\mathcal{I}\sim f_{\mathcal{I}}(I)}[log(1-D(T(I(\mathbf{x}),\theta),\vartheta))]\end{aligned} \tag{10}$$

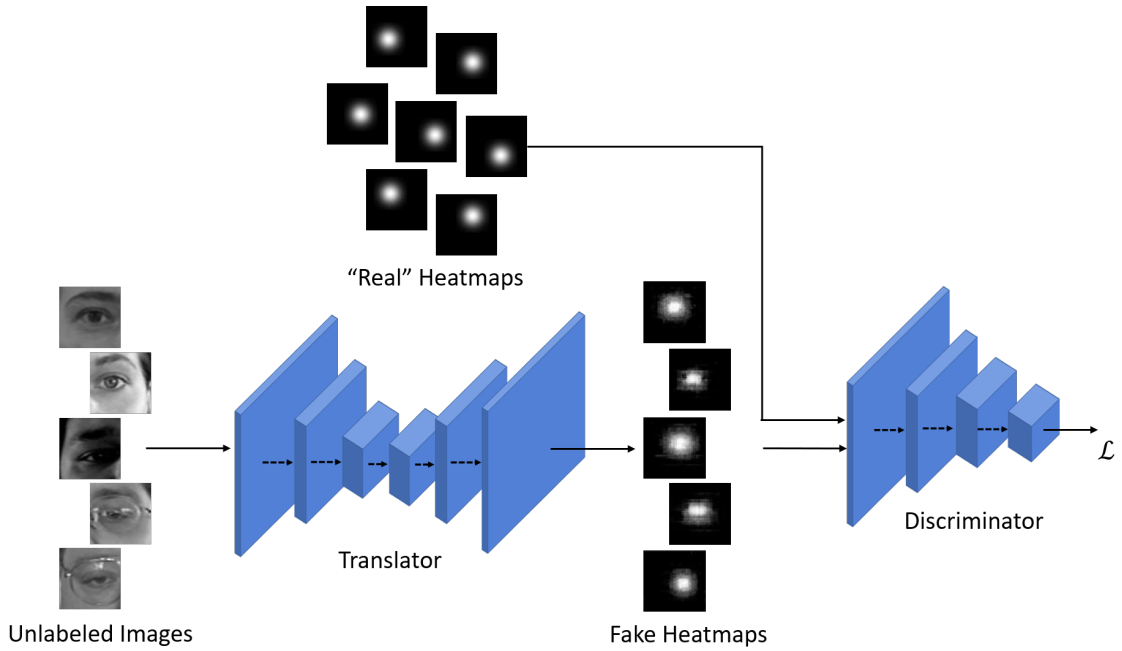

Figure 3. PupilTAN Deep Architecture

and solve the following $min - max$ optimization problem:

$$\min_{\theta} \max_{\vartheta} \mathcal{L}(\theta, \vartheta). \tag{11}$$

3.3. Network Architecture

As we can see from Figure 3 the Translational Adversarial Network is composed by the Translator and the Discriminator subnetworks. The Translator is a 3-stage encoder-decoder network that can fully exploit hierarchical feature representations for the transformation of the input images to their corresponding spatial idealizations, i.e., to heatmaps. The encoder comprises a pyramid structure of convolutional blocks followed by batch normalization, rectified linear and max-pooling layers in order to extract distinct geometrical information in different scales. In particular, the first layer consists of 128 channels and after each stage the channels are doubled in order the whole net to be able to learn the complex structures effectively. The decoder, on the other hand, uses transposed convolutions to up-sample the feature maps on different scales reducing the number of channels by a factor of two. All convolutions but the last are followed also by batch normalization and rectified linear layers. The final feature map is fed into a one-channel convolutional layer with a $tanh(.)$ activation function in order to aggregate better multi-scale information and obtain the final regression map.

The Discriminator consists of 4 fractionally-strided convolutional layers followed by batch normalization layers (all but the first) and leaky-rectified linear layers with negative slope set to 0.2. The first layer consists of 128 channels and after each stage the channels are doubled. Finally, a one-channel convolutional layer follows with a sigmoid activation function to form the output of the Discriminator. To prevent both networks from overfitting, we add Dropout layers with a rate of $p_{drop} = 0.5$ before the decoder part of the Translator and on the top of the Discriminator.

The proposed framework was trained for 300 epochs and batch size of 128 images. We use ADAM optimizer [28] with initial learning rate of 2×10^{-4} and momentum terms $\beta_1 = 0.5$ and $\beta_2 = 0.999$. To speed up the training process, we use a Nvidia GeForce GTX 1080 Ti GPU.

4. Experiments

4.1. Experimental Setup

Experiments were performed in two publicly available face databases in order to evaluate the performance of the proposed method. Specifically, the selected MUCT [29] and BioID [30] databases are among the most challenging and characteristic datasets and were widely used in previous eye-center localization techniques. The images where the face detector failed to detect the face due to extreme poses, were excluded for the experiments (2% for MUCT and 5.96% for BioID). The MUCT face database consists of 3755 low resolution (640×480 pixels) color images of frontal or near frontal faces, containing a wide variety of ages, ethnicities and light conditions. The images were acquired using five webcams from different positions, resulting in a pose variation. This, in combination with the occlusions from hair, glasses and reflections, increases extremely

its "difficulty" factor. The BioID database consists of 1521 grayscale images of 23 subjects taken at different times of the day in different positions with a low resolution camera (384×288 pixels). The size, position and pose of the faces varies. Furthermore, many subjects were wearing glasses, while in some instances the eyes were closed or hidden by strong reflections on glasses. Thus, BioID is regarded as one of the most challenging databases. For the purpose of eye center localization, 29 images that contain totally closed eyes were manually removed.

In order to evaluate the accuracy of the proposed method we adopted the normalized error, representing the worst eye center estimation of the two eyes. The normalized error (e) is defined as [31]:

$$e = \frac{\max\{||\hat{C}_L - C_L||_2,\ ||\hat{C}_R - C_R||_2\}}{||C_L - C_R||_2} \qquad (12)$$

where, $\hat{C}_L,\ \hat{C}_R$ are the estimations of the left and right eye center coordinates resulting from the application of the proposed method and $C_L,\ C_R$ are the manually labeled corresponding coordinates. The $||C_L - C_R||_2$ term represents the distance between the two real eye centers and is used as a normalization factor for the localization error. The accuracy of the algorithm is expressed by the ratio between the number of the eye center localizations that fall below the assigned error threshold and the total number of them. The threshold $e \leq 0.25$ represents the distance between the eye center and the eye corners, the $e \leq 0.1$ represents the range of the iris and the $e \leq 0.05$ represents the pupil area.

4.2. Experimental Results

The evaluation of the proposed method leads us to the conclusion of a robust and highly precise localization method. This method deals successfully with the most challenging circumstances including shadows, pose variations, occlusions by hair or strong reflections, out-of-plane rotations and presence of glasses (Figure 4).

4.2.1 Comparisons against state-of-the-art techniques

A comparison of the proposed method with the state-of-the-art methods is carried out and the results are presented on the following tables. All accuracies of the under comparison techniques are the published ones. To evaluate the accuracy of the proposed method, a 5-fold cross validation was adopted. This validation is performed by randomly dividing each dataset into 5 equal subsets and retaining every single subset for validation and the remaining ones for training. All the following tables provide supporting evidence of the superior performance of the proposed method in terms of accuracy. Table 1 contains the results obtained from the application of the proposed method and other relative works, on the MUCT database. It is evident that in the degraded images of this database, the proposed method achieved a significant improvement of 2.4% in performance over the best method for the fine accuracy level ($e \leq 0.05$).

Table 1. Accuracy vs. normalized error in the MUCT database

Method	Accuracy (%)		
	$e \leq 0.05$	$e \leq 0.1$	$e \leq 0.25$
PupilTAN	**97.15**	**99.32**	**100**
MFRST[2017] [7]	94.75	98.67	99.76
Skodras[2015] [6]	92.90	97.20	99.00
Timm[2011] [32]	78.60	94.90	98.60
Valenti[2008] [4]	63.10	76.70	94.10

Table 2. Accuracy vs. normalized error in the BioID database

Method	Accuracy (%)		
	$e \leq 0.05$	$e \leq 0.1$	$e \leq 0.25$
PupilTAN	**96.86**	99.71	**100**
Lee[2020] [18]	96.71	98.95	100
Choi[2020][17]	93.30	96.91	100
Xia[2019] [16]	94.40	**99.90**	100
Xiao[2018] [33]	94.35	98.75	99.80
Li[2018] [15]	85.60	95.90	99.50
Wang[2018] [34]	82.15	98.70	100
MFRST[2017] [7]	87.10	98.15	100
Anjith[2016] [35]	85.00	94.30	-
Cai[2015][36]	84.10	95.60	99.80

PupilTAN performance in the low resolution images of BioID face database is presented in Table 2 and compared with the state-of-the-art techniques. The proposed technique outperformed its rivals for the fine accuracy level ($e \leq 0.05$) while it achieved almost equal performance to the best method [16] (slightly lower by 0.19%) for the case of $e \leq 0.1$. The abovementioned results lead us to the conclusion of a significant improvement of the proposed method against the state-of-the-art methods.

4.2.2 Comparisons against the supervised counterpart

In order to highlight the effectiveness of the proposed adversarial training framework, in this paragraph we conduct comparisons with the same network trained in a supervised way, by performing image-to-image regression. Specifically, we trained the proposed encoder-decoder part using paired ROI images with the corresponding heatmaps derived from Eq. (1). As a loss function, we adopt the L_2 norm between the estimated and real heatmaps. For fair comparison, we use the same network architecture and the ADAM optimizer with the default training parameters [28]. Experiments performed on the BioID database and presented in Table 3, demonstrate that the proposed unsupervised adversarial framework outperforms the corresponding supervised by 1.64% for the case of $e \leq 0.05$.

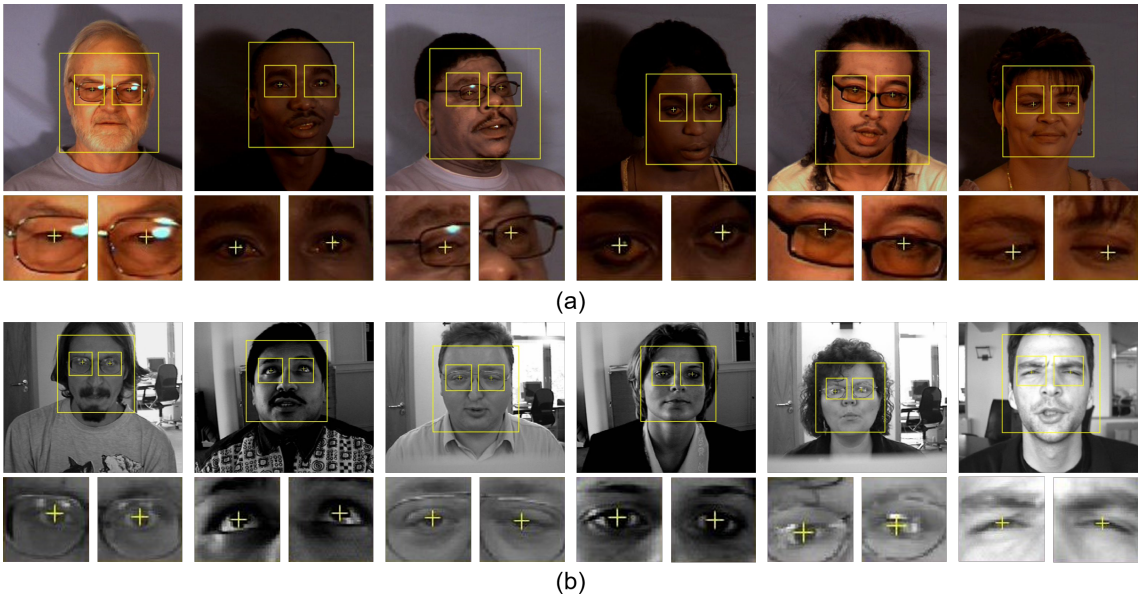

Figure 4. Precise eye center localization results on MUCT (a) and BioID (b) databases

Table 3. Accuracy comparisons between the proposed adversarial and the corresponding supervised frameworks

Method	Accuracy (%)		
	$e \leq 0.05$	$e \leq 0.1$	$e \leq 0.25$
Adversarial	**96.86**	**99.71**	**100**
Supervised	95.22	99.22	100

Table 4. PupilTAN performance for different architectures in the BioID database

N_{Par}	Time	Accuracy (%)		
		$e \leq 0.05$	$e \leq 0.1$	$e \leq 0.25$
3.18M	34ms	96.86	99.71	100
0.8M	19ms	96.29	99.71	100
0.2M	16ms	95.65	99.50	100

In general, unsupervised methods demonstrate inferior performance in comparison to their supervised counterparts, mostly due to the absence of the ground-truth data. However, results in Table 3, leads us to the conclusion that the proposed unsupervised framework succeeds in estimating the probability distribution of the Real heatmaps and mitigates the aforementioned issue. In this way, the adversarial network achieves a better generalization of the eye features and avoids to overfit to the training data resulting in enhanced localization accuracy. Note that, our intention is not to analyze ways to prevent from overfit and enhance the localization accuracy of the supervised method, but instead to highlight the advantage of using the proposed adversarial framework.

4.2.3 Comparisons against different architectures

In this section we analyze the impact of reducing the network parameters to the accuracy of the proposed method on the BioID database. Specifically, we decrease the size of the Translator by reducing the number of channels at each convolutional layer by a factor of two. Despite the accuracy decrease presented in Table 4, even the smallest model with $0.2M$ parameters still provides comparable accuracy to the other state-of-the-art methods. Note that, the performance after adding more parameters or layers saturates. Therefore, in terms of network complexity, PupilTAN is significantly reduced in comparison with other deep networks. Specifically, the architectures proposed in [18] and [17] contain $13.6M$ and $4.9M$ respectively only for the face detection and glasses removal networks, without considering the eye localization network. Moreover, the processing speed also increases when the network size decreases. For instance, the Translator requires only $16ms$ (Matlab implementation) to process every input face image for the smallest network.

5. Conclusions

In this paper, we introduced the PupilTAN, a few-shot adversarial training framework that performs image-to-heatmap translation for precise eye localization. In order to overcome the dependency of the labeled data, this framework aims to create artificial heatmaps, from a few ground-truth, that follow the same probability distribution of the real ones and train the Translator to accurately localize the eye centers. An extensive evaluation of the proposed method was performed on two publicly available databases

with low resolution images, containing many different cases of challenging conditions. Comparisons with existing methods demonstrated a significant improvement in accuracy over even supervised state-of-the-art techniques. Moreover, the robustness of the proposed deep network by significantly reducing the number of its parameters was highlighted.

Given the real-time performance achieved by the proposed method, we believe that this approach can be incorporated into low-cost eye trackers, where the localization accuracy is fundamental.

References

[1] A. Kar and P. Corcoran, "A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms," *IEEE Access*, vol. 5, pp. 16495–16519, August 2017.

[2] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2016, pp. 2176–2184.

[3] A. Gudi, X. Li, and J. van Gemert, "Efficiency in real-time webcam gaze tracking," in *European Conference on Computer Vision (ECCV) Workshops*. Springer, 2020, pp. 529–543.

[4] Valenti R. and T. Gevers, "Accurate eye center location through invariant isocentric patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 34, pp. 1785–1798, September 2012.

[5] G. Loy and A. Zelinsky, "Fast radial symmetry for detecting points of interest," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 25, pp. 959–973, August 2003.

[6] Skodras E. and N. Fakotakis, "Precise localization of eye centers in low resolution color images," *Image and Vision Computing*, vol. 12, pp. 537–543, August 2015.

[7] N. Poulopoulos and E. Z. Psarakis, "A new high precision eye center localization technique," in *International Conference on Image Processing (ICIP)*. IEEE, 2017, pp. 2806–2810.

[8] N. Poulopoulos and E. Z. Psarakis, "Real time eye localization and tracking," in *International Conference on Robotics in Alpe-Adria Danube Region (RAAD)*. Springer, 2018, pp. 560–571.

[9] M. Everingham and A. Zisserman, "Regression and classification approaches to eye localization in face images," in *International Conference on Automatic Face and Gesture Recognition*. IEEE, 2006, pp. 441–446.

[10] F. Samaria and S. Young, "Hmm-based architecture for face identification," *Image and Vision Computing*, vol. 12, pp. 537–543, August 1994.

[11] P. Campadelli, R. Lanzarotti, and G. Lipori, "Precise eye and mouth localization," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, pp. 359–377, August 2009.

[12] Z. Niu, S. Shan, S. Yan, X. Chen, and W. Gao, "2d cascaded adaboost for eye localization," in *International Conference on Pattern Recognition (ICPR)*. IEEE, 2006.

[13] I. S. Pandzic J. Ahlberg N. Markus, M. Frljak and R. Forchheimer, "Fast radial symmetry for detecting points of interest," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 47, pp. 578–587, 2014.

[14] W. Fuhl, "Pupilnet: convolutional neural networks for robust pupil detection," *arXiv*, vol. 1601.04902, January 2016.

[15] B. Li and H. Fu, "Real time eye detector with cascaded convolutional neural networks," *Applied Computational Intelligence and Soft Computing*, April 2018.

[16] Y. Xia, H. Yu, and F. Wang, "Accurate and robust eye center localization via fully convolutional networks," *IEEE/CAA Journal of Automatica Sinica*, vol. 6, pp. 1127–1138, September 2019.

[17] J.H. Choi, K.I. Lee, and B.C. Song, "Eye pupil localization algorithm using convolutional neural networks," *Multimedia Tools and Applications*, vol. 79, pp. 32563–32574, August 2020.

[18] KI. Lee, JH. Jeon, and BC. Song, "Deep learning based pupil center detection for fast and accurate eye tracking system," in *European Conference on Computer Vision (ECCV)*. Springer, 2020, pp. 1127–1138.

[19] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *International Conference on Neural Information Processing Systems (NIPS)*. IEEE, 2014, pp. 2672–2680.

[20] K. Regmi and A. Borji, "Cross-view image synthesis using conditional gans," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018, pp. 3501–3510.

[21] C. Lin, E. Yumer, O. Wang, E. Shechtman, and S. Lucey, "St-gan: Spatial transformer generative adversarial networks for image compositing," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2018, pp. 9455–9464.

[22] H. Zhang, T. Xu, H. Li, S. Zhang, X. Wang, X. Huang, and D. Metaxas, "Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks," in *International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 5909–5915.

[23] J. Zhu, T. Park, P. Isola, and A. Efros, "St-gan: Spatial transformer generative adversarial networks for image compositing," in *International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 2242–2251.

[24] A. Alotaibi, "Deep generative adversarial networks for image-to-image translation: A review," *Symmetry*, vol. 12, October 2020.

[25] P. Isola, J. Zhu, T. Zhou, and A. Efros, "Image-to-image translation with conditional adversarial networks," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 1125–1134.

[26] J. P. Robinson, Y. Li, N. Zhang, Y. Fu, and S. Tulyakov, "Laplace landmark localization," in *International Conference on Computer Vision (ICCV)*. IEEE, 2019, pp. 10102–10111.

[27] P. Viola and M. Jones, "Robust real-time face detection," *Int. Journal on Compututer Vision*, vol. 57, pp. 137–154, February 2004.

[28] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*. IEEE, 2015.

[29] S. Milborrow, J. Morkel, and F. Nicolls, "The muct landmarked face database," *Pattern recognition association of South Africa*, vol. 201, 2010.

[30] "The bioid face database," in *B. T. Research*, 2001.

[31] O. Jesorsky, K. J. Kirchbergand, and R. Frischholz, "Robust face detection using the hausdorff distance," *In Audio and Video Biometric Person Authentication*, pp. 90–95, 1992.

[32] F. Timm and E. Barth, "Accurate eye centre localization by means of gradients," in *International Conference on Computer Vision Theory and Applications (VISAPP)*, 2011, pp. 125–130.

[33] F. Xiao, K. Huang, Y. Qiu, and H. Shen, "Accurate iris center localization method using facial landmark, snakuscule, circle fitting and binary connected component," *Multimedia Tools and Applications*, vol. 77, pp. 25333–25353, February 2018.

[34] Z. Wang, H. Cai, and H. Liu, "Robust eye center localization based on an improved svr method," in *International Conference on Neural Information Processing*. Springer, 2018, pp. 623–634.

[35] G. Anjith and A. Routray, "Fast and accurate algorithm for eye localization for gaze tracking in low resolution images," *arXiv preprint arXiv:1605.05272*, 2016.

[36] H. Cai, H. Yu, C. Yao, S. Chen, and H. Liu, "Convolution-based means of gradient for fast eye center localization," in *International Conference on Machine Learning and Cybernetics*. IEEE, 2015.