Few-shot Gaze Estimation via Gaze Transfer

Nikolaos Poulopoulos[1][a], Emmanouil Z. Psarakis[1][b]
[1]*Department of Computer Engineering & Informatics, University of Patras, Greece*
{*npoul, psarakis*}*@ceid.upatras.gr*

Keywords: Gaze Estimation, Gaze Transfer, Gaze Tracking, Deep Neural Networks, Convolutional Neural Networks, Transfer Learning.

Abstract: Precise gaze estimation constitutes a challenging problem in many computer vision applications due to many limitations related to the great variability of human eye shapes, facial expressions and orientations as well as the illumination variations and the presence of occlusions. Nowadays, the increasing interest of deep neural networks requires a great amount of training data. However, the dependency on labeled data for the purpose of gaze estimation constitutes a significant issue because they are expensive to obtain and require dedicated hardware setup. To address these issues, we introduce a few-shot learning approach which exploits a large amount of unlabeled data to disentangle the gaze feature and train a gaze estimator using only few calibration samples. This is achieved by performing gaze transfer between image pairs that share similar eye appearance but different gaze information via the joint training of a gaze estimation and a gaze transfer network. Thus, the gaze estimation network learns to disentangle the gaze feature indirectly in order to perform precisely the gaze transfer task. Experiments on two publicly available datasets reveal promising results and enhanced accuracy against other few-shot gaze estimation methods.

1 INTRODUCTION

Eye gaze constitutes a revolutionary approach to interact without physical contact and provides a rich information of human intention, cognition and behavior (Eckstein et al., 2017). Nowadays, eye gaze is of growing interest providing a new input modality for various **human computer interaction** (HCI) applications like:

- **virtual reality** (Chen et al., 2020)
- **health care** and **analysis** (Huang et al., 2016)
- **self-driving cars** (Palazzi et al., 2019), etc.

Despite the active research in this field, the accuracy of such eye gaze systems has room for improvement and usually downgraded by many limitations. The main challenges are related to the wide variety of human eye shapes, the eye states (open or closed), the facial expressions and orientations, etc. Moreover, the presence of occlusions from hair and glasses, reflections and shadows as well as poor lighting and low image resolution further degrades the gaze estimation accuracy.

[a] https://orcid.org/0000-0002-8341-9805
[b] https://orcid.org/0000-0002-9627-0640

Obtaining high-quality data to train supervised gaze estimators constitutes an expensive and challenging task. This happens because the gaze direction can only be measured indirectly, using complicated hardware setups and geometry calculations. The limited labeled datasets usually lead supervised methods to overfit the training data. On the other hand, there is a plenty of unlabeled eye data available for free.

To address these limitations and become less dependent on labeled data, we introduce a few-shot learning approach which exploits a large amount of unlabeled data to disentangle the gaze feature and train a gaze estimator using only few calibration samples (e.g. 100). To achieve so, we perform gaze transfer between pairs of images that share similar eye appearance but different gaze information. To that end, a gaze transfer network and a gaze estimation network were trained jointly. The gaze estimation network aims to encode gaze information of the reference eye image, while the gaze transfer network aims to transfer the gaze of the input eye image to the one learned from the gaze estimation network. The main contributions of this work are summarized as follows:

- An **unsupervised** gaze representation learning approach, based on gaze transfer.

- An **extension** of image pairs selection with different head poses.
- **Enhanced** gaze estimation **accuracy** with only few calibration samples.

2 RELATED WORK

In this section, we review relevant works on gaze estimation and unsupervised representation learning. **Gaze estimation methods** can be divided into model-based and appearance-based methods. Model-based methods estimate gaze by fitting a geometric eye model to the eye image (Park et al., 2018), (Wang and Ji, 2018) and rely on accurately detected facial features (e.g. eye corners or eye centers) (Poulopoulos and Psarakis, 2022a), (Poulopoulos and Psarakis, 2022b). However, the accuracy of these methods highly depends on the image resolution and the illumination thus resulting into degraded performance in real-world scenarios. Appearance-based methods directly regress the gaze vector from the eye images and nowadays outperform model-based methods in terms of accuracy (Zhang et al., 2019). While early works assumed a fixed head pose (Lu et al., 2014), recent works allow an unconstrained head movement in relation to the camera (Kellnhofer et al., 2019). Deep CNNs have also achieved several improvements over the last years. Krafka et al. (Krafka et al., 2016) indicated that a multi-region CNN considering the eye regions and the face as inputs can benefit gaze estimation performance. Zhang et al. (Zhang et al., 2017) introduced a CNN with a spatial weights mechanism in order to enhance the gaze-related information. Cheng at al. (Cheng et al., 2018) exploited the asymmetric performance of the left and right eyes using an evaluation network in order to improve the gaze accuracy. A data augmentation approach for improving the gaze estimation has been proposed by Zheng et al. (Zheng et al., 2020). Although the aforementioned methods perform well on within dataset evaluations, they lack of accuracy when tested on new data. This happens because they strongly depend on the amount and diversity of training data which are limited due to the difficulty to collect accurate 3D gaze annotations. Recently, there is an increasing interest in collecting synthetic data to overcome this limitation (Wood et al., 2015),(Wood et al., 2016), but the domain gap between them and the real ones still remains a crucial issue.

Unsupervised representation learning aims to learn specific features from unlabeled images. Such methods were proposed to solve object detection (Crawford and Pineau, 2019) and localization (Poulopoulos et al., 2021), image classification (Caron et al., 2018) and semantic segmentation problems (Moriya et al., 2018). Yu et al. (Yu and Odobez, 2020) were the first to learn unsupervised gaze representation via gaze redirection. They used the gaze representation difference of paired images with similar head pose to feed a gaze redirection network. Cross-encoder proposed in (Sun and Chen, 2021) aimed to disentangle the gaze feature from the eye related features by reconstructing pairs of images with switched latent features. Gideon et al. (Gideon and Stent, 2021) extended this work for the case of multi-view face video sequences. Despite the growing interest, unsupervised gaze representation learning remains challenging due to the difficulty to disentangle the gaze feature without the annotations. Our work was inspired by the work proposed in (Yu and Odobez, 2020) and tried to overcome the aforementioned challenges by learning the gaze-related features via a joint training of a gaze transfer and a gaze estimation networks with unlabeled pairs of images. We believe that forcing the gaze estimation network to learn directly the gaze feature from the reference images instead of the gaze angle differences (Yu and Odobez, 2020) can benefit gaze estimation performance. Moreover, we showcase that importing the head pose information into both networks permit us to overcome the constraint of similar head poses between the training pairs.

3 THE PROPOSED METHOD

In this section, we are going to give a detailed description of the proposed framework, as well as, the network details and training options.

3.1 Overview

The main idea of the proposed unsupervised gaze representation approach is shown in Figure 1. As it can be seen, the proposed framework is composed by:

- a **gaze estimation** network $G_e(.,.;\theta)$ and
- a **gaze transfer** network $G_t(.,.;\phi)$

with θ, ϕ denoting their parameters. Both networks are trained jointly using pairs of unlabeled images. Specifically, we consider that:

- **input** image $\mathbf{i}_{in}$ and **target** image $\mathbf{i}_t$, with $\mathbf{i}$ denoting the column-wise vectorized version of image I, share:
  - **similar** eye appearance but
  - **different** gaze direction, while

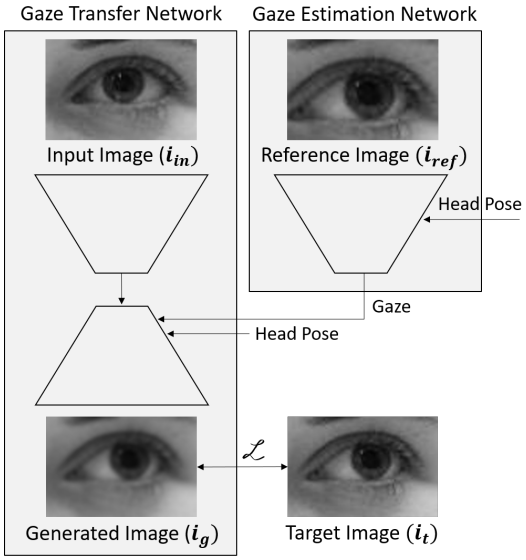

Figure 1: Proposed unsupervised gaze representation learning. Input and Target images share similar eye appearance but different gaze direction. Gaze transfer network transfers the input gaze to the estimated gaze from the gaze estimation network.

- **reference** image $\mathbf{i}_{ref}$ results from an **unknown** transformation which, however, **preserves** the gaze information of the input image $\mathbf{i}_{in}$.

The aim of the whole framework is **to force** the gaze estimation network **to learn** the gaze of the reference image in order to transfer the gaze of the input image via the gaze transfer network. The generated image $\mathbf{i}_g$ has to be as close as possible to the target one, that is the $\mathbf{i}_t$. Note that the image pairs should be taken from the same person, but contrary to (Yu and Odobez, 2020), can share different head poses, as the pose information is directly imported into both networks.

Having completed our presentation of the proposed framework for the gaze problem, we are going in the next subsection to present our data driven unsupervised approach.

3.2 Unsupervised Gaze Representation

To this end, let us consider the following set of **training paired** grayscale images and the head pose vector, consisted of the polar and azimuthal angles:

$$\mathcal{S}_{\mathbf{i}} = \left\{ \mathbf{i}_{in_k},\ \mathbf{i}_{tk},\ \mathbf{h}_{tk} \right\}_{k=1}^{K} \tag{1}$$

with each member of this set constituting a realization of the random variable I whose multivariate pdf, $f_I(\mathbf{i})$ is known, and $\mathbf{i}_{in}$, $\mathbf{i}_t$ represent realizations of the input and target images respectively with the last, as it was mentioned in the previous subsection, having the same eye appearence but different gaze direction, and $\mathbf{h}_t$ the head pose vector of the target one.
In addition, we consider that the **reference images** $\mathbf{i}_{ref}$ are derived from the application of a **gaze preserving** transform, that is it is restricted to be a translation and/or a scaling, to the target images $\mathbf{i}_t$, i.e.:

$$\mathbf{i}_{ref} = T(\mathbf{i}_t). \tag{2}$$

Note that under the above mentioned tranform the head pose vector $\mathbf{h}_t$ is also preserved. During the training phase, given the head pose vector $\mathbf{h}$ of the target image $\mathbf{i}_t$ the goal of $G_e(\mathbf{i},\ \mathbf{h};\theta)$ net is to learn the distribution of the gaze feature $\mathbf{g}_{ref}$, that is:

$$\mathbf{g}_{ref_{\theta,k}} = G_e(\mathbf{i}_{ref_k}, \mathbf{h}_{t_k}; \theta). \tag{3}$$

Thus, after its training, each value of its output $\mathbf{g}_{ref_{\theta^\star,k}}$ will constitute a realization of this random variable.
On the other hand, the goal of the $G_t(\mathbf{i};\phi)$ is to transfer the gaze of the $\mathbf{i}_{in_k}$ according to the $\mathbf{g}_{ref_{\theta,k}}$, i.e.:

$$\mathbf{i}_{g_k}(\theta,\phi) = G_t(\mathbf{i}_{in_k}, G_e(\mathbf{i}_{ref_k}, \mathbf{h}_{t_k}; \theta), \mathbf{h}_{t_k};\ \phi) \tag{4}$$

or, by using Eq. (3), the above equation can be equivalently rewritten as:

$$\mathbf{i}_{g_k}(\theta,\phi) = Gt(\mathbf{i}_{in_k},\ \mathbf{g}_{ref_{\theta,k}},\ \mathbf{h}_{t_k};\ \phi). \tag{5}$$

It is clear that we would like after the training of this net, its output to reproduce the realizations of $\mathbf{i}_t$.
In order to achieve it, both networks are trained jointly by minimizing the following loss function:

$$\mathcal{L}(\theta,\ \phi) = \mathbb{E}_{I\sim f_I}\left[||\mathbf{i}_t - \mathbf{i}_g(\theta,\ \phi)||_2^2 \right]. \tag{6}$$

In this way, the gaze estimation network is trained indirectly to disentangle the gaze feature of the reference image in order the gaze transfer network to generate an image close to the target one.

3.3 Few-shot Gaze Estimation

During unsupervised training, gaze estimation network learns a gaze representation from unlabeled images. In order to map this representation to the real gaze angles and estimate the gaze in the camera coordinate system, we follow a two-step procedure. Firstly, we add a MLP layer at the end of the gaze estimation network and train only this layer using a few calibration samples. Then, in order to further adapt to the calibration samples, we fine-tune all the weights of the network using these samples. During this process, the network weights were initialized from the preceded unsupervised training and retrained for few more iterations in order to better fit to the data. Note that the second step is crucial for the accuracy of the estimator.

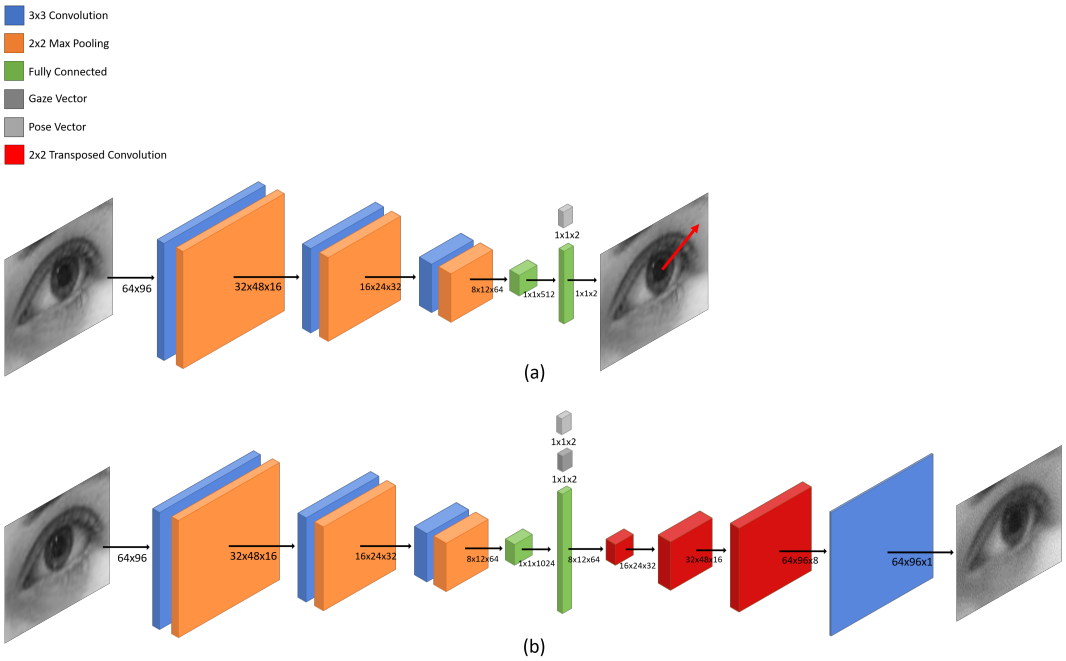

Figure 2: Architecture details of the gaze estimation (a) and gaze transfer (b) networks.

3.4 Network Details

The proposed architecture depicted in Figure 2, consists, as it was already mentioned, of:

- the **gaze estimation** and
- the **gaze transfer**

networks.

The **gaze estimation** network (Figure 2(a)) consists of three convolutional layers each one followed by a rectified linear and a max-pooling layer in order to extract features in different scales. In particular, the first convolutional layer consists of 16 channels and after each stage the number of channels are doubled. The last layer is followed by two fully connected layers with 512 and 2 outputs respectively. Moreover, the head pose is concatenated with the first fully connected layer. Note that the output of the network, similarly to (Yu and Odobez, 2020), is set to be of dimension 2, in order to avoid encoding eye related features except from the gaze.

The **gaze transfer** network (Figure 2(b)) is a three-stage encoder-decoder network. The encoder comprises a pyramid structure of three convolutional blocks followed by rectified linear and max-pooling layers. The first convolutional layer consists of 16 channels and after each stage the channels are doubled. On the other hand, the decoder 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 and rectified linear layers. The bottleneck between the encoder and decoder consists of a fully connected layer with dimension of 1024, where the gaze and head pose vectors are concatenated. 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 generated image.

3.5 Implementation Details

Every face image was cropped according to the detected facial features (Kartynnik et al., 2019) in order to derive the corresponding eye image and then transformed to grayscale and resized to the size of 64x96 pixels. All experiments were conducted using only the right eye images. The gaze feature is highly correlated with the eye-related features (Sun and Chen, 2021). Thus, in order to disentangle the gaze feature, we apply a gaze-preserving transformation to the reference images, similarly to (Yu and Odobez, 2020). Specifically, the applied random translation and scaling transformations affect the eye feature positions but not the gaze direction. This transformation improves significantly the accuracy of the gaze estimator, as shown in the next section. The proposed framework was trained for 150 epochs with a batch size of 256 images, using ADAM optimizer (Kingma and Ba, 2015) with default parameters. To speed up the training process, we use a Nvidia GeForce GTX 1080 Ti GPU.

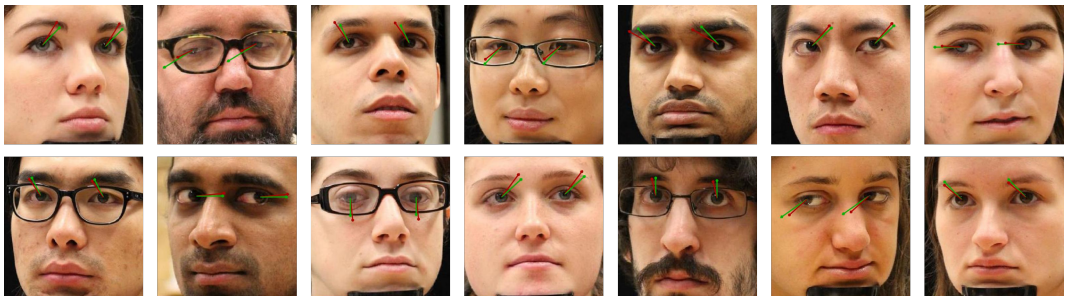

Figure 3: Sample estimates (red) and ground-truth (green) after the application of the proposed method on Columbia dataset.

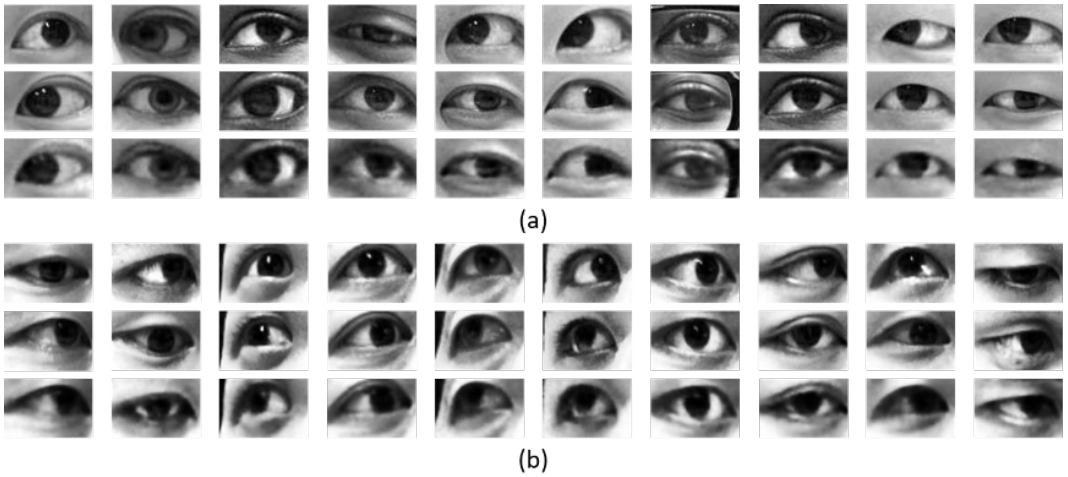

Figure 4: Gaze transfer results on Columbia (a) and UTMultiview (b) datasets. The first row corresponds to the input images, while the second and third rows to the target and generated images respectively.

4 EXPERIMENTS

4.1 Experimental Setup

Datasets: Experiments were performed on two publicly available gaze databases in order to evaluate the performance of the proposed training scheme. Specifically, Columbia Gaze (Smith et al., 2013) consists of 5880 high resolution images from 56 people over 5 head poses and 21 gaze directions per head pose with a great variety of ages and ethnicities. UTMultiview (Sugano et al., 2014) consists of 64000 images of 50 people with 160 gaze directions using eight (8) cameras. Images contain a wide variety of photometric distortions and shadows.

Validation Settings: Exploiting the division of the selected Columbia and UTMultivew datasets into 5 and 8 head poses, we performed 5-fold and 8-fold within-dataset evaluation respectively. In each fold, the training data were used for unsupervised learning of the entire framework and then, the selected 100 random annotated samples for few-shot fine-tunning of the gaze estimation network. Note that the training pairs were selected randomly with the only constraint to be from the same person (similar eye appearance). The remaining test data were used only for validation. All experiments were performed 5 times and the reported results are the mean errors.

Evaluation Metric: In order to evaluate the accuracy of the proposed method we adopted as a metric the angular error in degrees. Let $\hat{\mathbf{g}}$ be the 3-dimensional predicted gaze vector with respect to the camera coordinate system, after the fine tuning of the whole net, and $\mathbf{g}$ the ground-truth. Then, the angular error is defined as follows:

$$\Delta\phi_{gaze} = \frac{180}{\pi} \arccos\left(< \frac{\mathbf{g}}{||\mathbf{g}||_2}, \frac{\hat{\mathbf{g}}}{||\hat{\mathbf{g}}||_2} > \right) \quad (7)$$

where $< ., . >$, $||\mathbf{x}||_2$ denote the inner product operator and the l_2 norm of vector $\mathbf{x}$ respectively.

4.2 Experimental Results

The qualitative evaluation of the proposed method demonstrates that it is highly accurate and robust. Figure 3 depicts random results of the proposed gaze

Table 1: Mean angular error of 100-shot gaze estimation on Columbia and UTMultiview datasets.

Method	**Dataset**	
	Columbia	UTMultiview
Proposed	**6.1**	**7.1**
Cross-Encoder (Sun and Chen, 2021)	6.4	7.4
Yu2020 (Yu and Odobez, 2020)	7.15	7.88
SimCLR (Chen et al., 2020)	7.2	12.1
BYOL (Grill et al., 2020)	9.9	14.4

Table 2: Mean angular error of 100-shot gaze estimation when trained on UTMultiview and tested Columbia dataset.

Method	**Angular error**
Proposed	8.5
Cross-Encoder (Sun and Chen, 2021)	**7.48**
Yu2020 (Yu and Odobez, 2020)	8.82

Table 3: Accuracy decrease on Columbia dataset when removing certain parts of the proposed framework.

Angular error	**MLP**	**Fine tune**	**Head Pose**
6.1	✓	✓	✓
7.3	✓	✓	
9.1	✓		

estimation network applied to Columbia database. For better visualization, the estimated gaze angle from the right eye is also displayed on the left eye. As it can be seen, the proposed gaze estimator achieves accurate results even under extreme head poses. Moreover, the quantitative evaluation of the learned gaze estimator demonstrates enhanced accuracy over other few-shot gaze estimation methods. The evaluation was performed under both within-dataset and cross-dataset settings.

4.2.1 Within-Dataset Evaluation

The accuracy of the proposed training scheme performing within-dataset experiments on Columbia and UTMultiview was compared against other few-shot gaze estimation methods. Table 1 presents the comparison results using 100 calibrations samples. Note that there are limited few-shot gaze estimation methods available for comparison in the literature. The proposed learning framework demonstrates enhanced accuracy over the rest of the methods both on Columbia and UTMultivew datasets. Compared to Yu (Yu and Odobez, 2020) method, it seems that forcing the gaze estimation network to learn directly the gaze feature from the reference images instead of the gaze angle differences can benefit gaze estimation performance. It is worth mentioning that all the accuracies from the compared methods are the published ones. Moreover, the accuracies from contrastive learning methods SimCLR (Chen et al., 2020) and BYOL (Grill et al., 2020) derive from (Sun and Chen, 2021).

4.2.2 Cross-Dataset Evaluation

In order to investigate the performance on totally unseen images, a cross-dataset evaluation was performed using the UTMultiview dataset for training and the Columbia dataset for testing. Table 2 presents the results from the proposed method as well as from other few-shot gaze estimation methods under the same training and testing format. The proposed method performs better compared to Yu (Yu and Odobez, 2020) method, however, it lacks of accuracy compared to Cross-Encoder (Sun and Chen, 2021). This accuracy decrease may results from the great diversity between the head pose angles of Columbia and UTMultiview datasets.

4.2.3 Gaze Transfer

The proposed framework aims to learn an unsupervised gaze representation indirectly via the joint training of two networks, a gaze estimation and a gaze transfer network. Although this work emphasizes on the gaze estimation performance, it worth mentioning that a highly precise gaze transfer network has also been trained in unsupervised way. Figure 4 illustrates the gaze transfer results based on image pairs from Columbia and UTMultiview databases. The first row corresponds to the input images, while the second and third rows to the target and generated images respectively. As can be seen, the network achieves precise gaze transfer between the image pairs.

4.3 Ablation Study

In order to investigate the contribution of each part of the propose framework to the final accuracy we perform experiments on Columbia database. Specifically, we studied the impact of fine-tuning the gaze estimation network using the calibration samples, as well as, the impact of importing the head pose information to the final performance. Results presented on Table 3 demonstrate the importance of these parts. Specifically, the head pose information increases accuracy by 1.2^o, while the fine-tuning of the network adds 1.8^o more accuracy increase.

Finally, we studied the impact of the applied gaze-preserving transformation of Eq. (2) to the reference images. Results showed an accuracy decrease of 1.6^o (from 6.1^o to 7.7^o) after removing this step, revealing that this step is crucial in order to disentangle the gaze feature from the eye-related features.

5 CONCLUSIONS

In this paper a few-shot gaze estimation method was introduced. In order to overcome the dependency of the labeled data, the proposed framework aimed to learn an unsupervised gaze representation via the joint training of a gaze transfer and a gaze estimation network. Only few calibration samples were enough to fine-tune the gaze estimation network with promising accuracy results. Extensive evaluation of the proposed method was performed on two publicly available databases. A comparison with existing few-shot gaze estimation methods demonstrated a significant improvement in accuracy in within-dataset experiments. Also, the benefits of every individual step of the proposed framework to the achieved performance were highlighted. The validity of this work makes us believe that this approach can be used as a pretraining process in order to exploit the great amount of the existing unlabeled data and become less dependent from the labeled ones.

ACKNOWLEDGEMENTS

This research has been co-financed by the European Union and Greek national funds through the Operational Program Competitiveness, Entrepreneurship and Innovation, under the call RESEARCH – CREATE – INNOVATE (project SignGuide, code: T2EDK - 00982)

REFERENCES

Caron, M., Bojanowski, P., Joulin, A., and Douze, M. (2018). Deep clustering for unsupervised learning of visual features. In *European Conference on Computer Vision (ECCV)*, pages 132–149.

Chen, M., Jin, Y., Goodall, T., Yu, X., and Bovik, A. C. (2020). Study of 3d virtual reality picture quality. *IEEE Journal of Selected Topics in Signal Processing*, 14:89–102.

Cheng, Y., Lu, F., and Zhang, X. (2018). Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *European Conference on Computer Vision (ECCV)*, pages 100–115.

Crawford, E. and Pineau, J. (2019). Spatially invariant unsupervised object detection with convolutional neural networks. In *AAAI Conference on Artificial Intelligence*, pages 3412–3420.

Eckstein, K. M., Guerra-Carrillo, B., Miller Singley, A. T., and Bunge, A. S. (2017). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development? *Developmental Cognitive Neuroscience*, 25:69–91.

Gideon, J., S. S. and Stent, S. (2021). Unsupervised multi-view gaze representation learning. In *International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 5001–5009.

Grill, J., Strub, F., Altch?, F., Tallec, C., Richemond, P., Buchatskaya, E., Doersch, C., Avila Pires, B., Guo, Z., Gheshlaghi Azar, M., and Piot, B. (2020). Bootstrap your own latent-a new approach to self-supervised learning. In *Advances in Neural Information Processing Systems*, pages 21271–21284.

Huang, M. X., Li, J., Ngai, G., and Leong, H. V. (2016). Stressclick: Sensing stress from gaze-click patterns. In *24th ACM International Conference on Multimedia*, pages 1395–1404.

Kartynnik, Y., Ablavatski, A., Grishchenko, I., and Grundmann, M. (2019). Real-time facial surface geometry from monocular video on mobile gpus. In *International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*.

Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., and Torralba, A. (2019). Gaze360: Physically unconstrained gaze estimation in the wild. In *International Conference on Computer Vision (ICCV)*.

Kingma, D. and Ba, J. (2015). Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Krafka, K., Khosla, A., Kellnhofer, P., Kannan, H., Bhandarkar, S., Matusik, W., and Torralba, A. (2016). Eye tracking for everyone. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2176–2184.

Lu, F., Sugano, Y., Okabe, T., and Y., S. (2014). Adaptive linear regression for appearance-based gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 36:2033–2046.

Moriya, T., Roth, H., Nakamura, S., Oda, H., Nagara, K., Oda, M., and Mori, K. (2018). Unsupervised segmen-

tation of 3d medical images based on clustering and deep representation learning. In *Biomedical Applications in Molecular, Structural, and Functional Imaging*, pages 483–489.

Palazzi, A., Abati, D., Calderara, S., Solera, F., and Cucchiara, R. (2019). Predicting the drivers focus of attention: The dr(eye)ve project. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 41:1720–1733.

Park, S., Zhang, X., Bulling, A., and Hilliges, O. (2018). Learning to find eye region landmarks for remote gaze estimation in unconstrained settings. In *ACM Symposium on Eye Tracking Research amp; Applications (ETRA)*, pages 1–10.

Poulopoulos, N. and Psarakis, E. (2022a). Deeppupil net: Deep residual network for precise pupil center localization. In *International Conference on Computer Vision Theory and Applications (VISAPP)*, pages 297–304.

Poulopoulos, N. and Psarakis, E. (2022b). A real-time high precision eye center localizer. *Journal of Real-Time Image Processing*, 19:475–486.

Poulopoulos, N., Psarakis, E., and Kosmopoulos, D. (2021). Pupiltan: A few-shot adversarial pupil localizer. In *International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3128–3136.

Smith, B., Yin, Q., Feiner, S., and Nayar, S. (2013). Gaze locking: passive eye contact detection for human-object interaction. In *ACM Symposium on User Interface Software and Technology (UIST)*, pages 271–280.

Sugano, Y., Matsushita, Y., and Sato, Y. (2014). Learning-by-synthesis for appearance-based 3d gaze estimation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1821–1828.

Sun, Y., Z. J. S. S. and Chen, X. (2021). Cross-encoder for unsupervised gaze representation learning. In *International Conference on Computer Vision (ICCV)*, pages 3702–3711.

Wang, K. and Ji, Q. (2018). 3d gaze estimation without explicit personal calibration. *Developmental Cognitive Neuroscience*, 79:216–227.

Wood, E., Baltrusaitis, T., Morency, L., Robinson, P., and Bulling, A. (2016). Learning an appearance-based gaze estimator from one million synthesized images. In *ACM Symposium on Eye Tracking Research Applications (ETRA)*, pages 131–138.

Wood, E., Baltrusaitis, T., Zhang, X., Sugano, Y., Robinson, P., and Bulling, A. (2015). Rendering of eyes for eye-shape registration and gaze estimation. In *International Conference on Computer Vision (ICCV)*, pages 3756–3764.

Yu, Y. and Odobez, J. (2020). Unsupervised representation learning for gaze estimation. In *International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7314–7324.

Zhang, X., Sugano, Y., and Bulling, A. (2019). Evaluation of appearance-based methods and implications for gaze-based applications. In *CHI conference on human factors in computing systems*, pages 1–13.

Zhang, X., Sugano, Y., Fritz, M., and Bulling, A. (2017). Its written all over your face: Full-face appearance-based gaze estimation. In *International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 51–60.

Zheng, Y. ad Park, S., Zhang, X., De Mello, S., and Hilliges, O. (2020). Self-learning transformations for improving gaze and head redirection. In *Advances in Neural Information Processing Systems (NIPS)*, pages 13127–13138.