2023 IEEE 26th International Conference on
Intelligent Transportation Systems (ITSC)
24-28 September 2023. Bilbao, Bizkaia, Spain

GSA-Gaze: Generative Self-adversarial Learning for Domain Generalized Driver Gaze Estimation

Hongcheng Han[1], Zhiqiang Tian[2], Yuying Liu[1], Shengpeng Li[1], Dong Zhang[1,3], Shaoyi Du[1,*], *Member, IEEE*

2023 IEEE 26th International Conference on Intelligent Transportation Systems (ITSC) | 979-8-3503-9946-2/23/$31.00 ©2023 IEEE | DOI: 10.1109/ITSC57777.2023.10421891

***Abstract*— Estimating driver gaze accurately is critical for the human-machine cooperative driving, but the significant facial appearance diversions caused by background, illumination, personal characteristics, etc. pose a challenge to the generalizability of gaze estimation models. In this paper, we propose the generative self-adversarial learning mechanism for generalized gaze estimation that aims to learn general gaze features while eliminating sample-specific features and preventing cross-domain feature over-fitting. Firstly, to reduce information redundancy, the feature encoder is designed based on pyramid-grouped convolution to extract a sparse feature representation from the facial appearance. Secondly, the gaze regression module supervises the model to learn as many gaze-relevant features as possible. Thirdly, the adversarial image reconstruction task prompts the model to eliminate the domain-specific features. The adversarial learning of the gaze regression and the image reconstruction tasks guides the model to learn only general gaze features across domains, preventing cross-domain feature over-fitting, enhancing the domain generalization capability. The results of cross-domain testing of four active gaze datasets prove the effectiveness of the proposed method. The code is available at https://github.com/HongchengHan/GSA-Gaze**

I. INTRODUCTION

Human-machine cooperative driving (HMCD) is a promising direction for the intelligent transportation system[1]. In order to achieve effective cooperative control, the vehicle must be capable of accurately comprehending the driver's intentions through various behaviors. Gaze is a crucial nonverbal communication cue[2], through which the driver's attention focus can be extracted in real-time to discern his/her driving intention, providing a foundation for cooperative control. Accurate driver gaze estimation methods are of significant importance for the advancement of HMCD vehicles.

Appearance-based gaze estimation methods are widely utilized nowadays due to their ease of implementation. To precisely estimate gaze from appearance in appearance-based methods, a robust feature extractor is essential for obtaining high-quality gaze feature representation from facial or ocular images. With the rapid advancement of deep learning in recent years, visual feature extraction algorithms are becoming increasingly potent, the appearance-based gaze estimation methods have been enhanced by the integration of deep neural networks[3]. Despite the impressive performance of deep-learning-based gaze estimation methods on various appearance-based gaze datasets, there still exist practical challenges that need to be addressed. The deep neural networks possess formidable function fitting capabilities and are capable of constructing intricate gaze feature representations, yet their learning is solely driven by the training data. When the test data and the training data differ in terms of background, illumination, and personal appearance, i.e., they are in different domains, deep-learning-based methods exhibit a significant performance drop in cross-domain testing.

To tackle cross-domain problems, a potential solution is to augment the training data with more diverse samples. Nevertheless, in vehicle scenes, due to privacy concerns and the high cost of labeling, developers are unable to collect customer data on a large scale. As a result, models can only be trained on limited data, making it difficult to cover all application scenarios. Another solution is domain adaption[4], by modeling the dissimilarity between the target and source domains, domain adaption enhances the performance of the model trained on the source domain in target domain testing. The methods based on domain adaption require only a certain number of unlabeled samples from target domains, reducing the cost of data acquisition and labeling. However, they still have certain limitations. Firstly, they still require samples from target domain, which fails to address the fundamental issue of data acquisition. Secondly, each target domain necessitates a distinct domain discriminator, thereby mandating separate model training for generalization from source to diverse target domains. These challenges render the deployment of gaze estimation algorithms in diverse environments a complex task.

With the above consideration, in this paper, we propose the generative self-adversarial learning mechanism (GSAL) to address the domain generalization problem in gaze estimation. The primary contributions of our work can be summarized as follows:

- We propose a novel method for generalized gaze estimation by adversarial learning of gaze regression and facial appearance reconstruction tasks, which focuses on general gaze features while eliminating domain-specific features, enhancing the generalizability.
- We propose the pyramid-grouped convolution for feature encoder network to enhance the sparsity of extracted features, thereby addressing information redun-

This work was supported by the National Key Research and Development Program of China under Grant No. 2020AAA0108100.

*Corresponding author: Shaoyi Du. dushaoyi@xjtu.edu.cn

[1]Hongcheng Han, Yuying Liu, Shengpeng Li and Shaoyi Du are with National Key Laboratory of Human-Machine Hybrid Augmented Intelligence, National Engineering Research Center for Visual Information and Applications, and Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an 710049, China.

[2]Zhiqiang Tian is with School of Software Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China.

[3]Dong Zhang is with School of Automation Science and Engineering, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China.

dancy issues encountered in general gaze pattern learning.

- Our method achieves domain generalization gaze estimation across multiple target domains without any prior information of the target domain data and requires only one training session on the source domain for all target domains.

II. RELATED WORK

The advancement of gaze estimation algorithms is propelled by the development of deep learning, which enhances precision but faces limitations in domain generalization. Although some domain adaptation methods have been proposed to address this issue, they often require target samples and are not universally applicable.

Gaze Estimation: Some researchers have adapted algorithms from fundamental computer vision tasks to enhance the accuracy of gaze estimation, e.g. Chen et al.[5] leverage dilated convolution for gaze estimation. Some works exploit the powerful feature extraction capabilities of deep neural networks to explore novel challenges in the field of gaze estimation. Cheng et al.[6] explore the two-eye asymmetry. Bao et al.[7] leverage face and eye images to estimate the gaze point. Zheng et al.[8] propose a gaze/head redirection network and employ synthesized images for data augmentation. These works have made significant contributions to enhancing the baseline of gaze estimation, however, the issue of cross-domain generalization remains unresolved.

Domain Adaption: Domain adaptive methods aim to address the performance drop problem of machine learning models in cross-domain testing. Yaroslav et al.[9] propose domain adaption neural network to model the dissimilarity by constructing the domain discriminator. Wang et al.[10] and Kellnhoder et al.[11] propose the utilization of adversarial learning to align features between the source and target domains. Liu et al.[12] propose an assemblage of networks that synergistically learn under the tutelage of anomalies. These methods rely on data from the target domain, which may not always be user-friendly. Achieving domain generalization without utilizing samples from the target domain remains a challenging task.

III. APPROACH

A. Purpose of Generative Self-adversarial Learning

The data-driven deep learning models extract not only general gaze features that are applicable across domains, but also the unique characteristics of the source domain, such as background, illumination, and personal facial appearance, as Fig. 1(a) and (b) show. This phenomenon is referred to as "cross-domain feature over-fitting". How to extract gaze features that are more generalizable and fewer domain-specific features is crucial for generalized gaze estimation.

The purpose of the proposed generative self-adversarial learning (GSAL) is to extract the common gaze features shared by source and target domains while eliminating domain-specific features unique to the source domain, addressing the cross-domain feature over-fitting problem. As

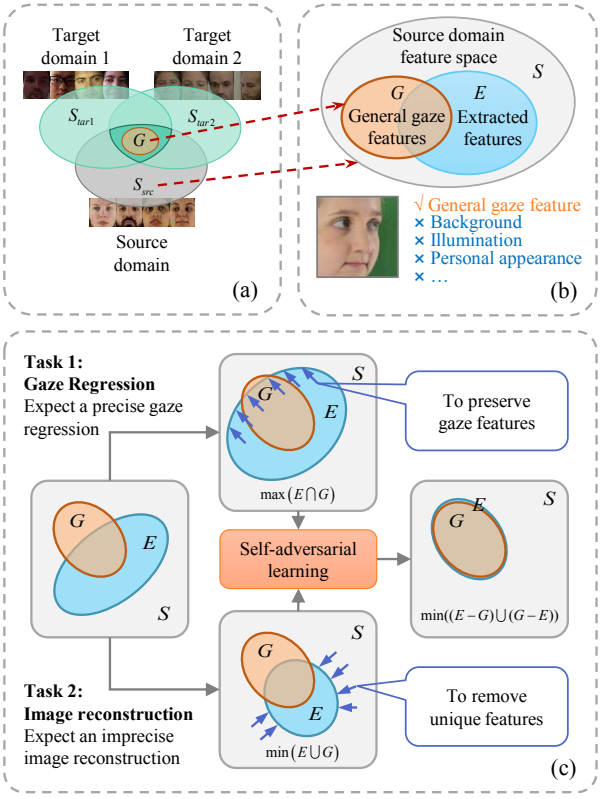


Fig. 1: Purpose of generative self-adversarial learning. (a) Cross-domain general gaze feature. S_{src}, S_{tar1}, S_{tar2} are the feature spaces of the source domain and two target domains, G is the general gaze feature for all domains. (b) Relationship of the general gaze feature and the extracted feature when training on the source domain. S is the whole feature space of the source domain, E indicates the extracted feature, G refers to the general gaze feature. (c) Principle of generative self-adversarial learning. To enhance the cross-domain generalization capability of the model, E is expected to match G.

Fig. 1(c) shows, the adversarial learning of the model is driven by two tasks. First, the gaze regression task, which expects a precise gaze regression, encouraging the model to learn more gaze-related features, expressed as

$$\max_{E(\boldsymbol{\theta}) \in S,\, G \in S} (E(\boldsymbol{\theta}) \cap G)\,, \tag{1}$$

where $\max(\cdot)$ means to maximize the feature set, E is the set of the extracted features, G is the set of the general gaze features, $\boldsymbol{\theta}$ refers to the learnable parameters of the feature extractor. Second, the image reconstruction task, which requires to reconstruct the original input image from the extracted features, the more unique features are extracted, the more similar the reconstructed image will be to its origin. However, we anticipate an imprecise reconstruction, and the more closely the reconstructed image resembles the original, the higher penalty we impose on the feature extractor. Therefore, in the image reconstruction task, the

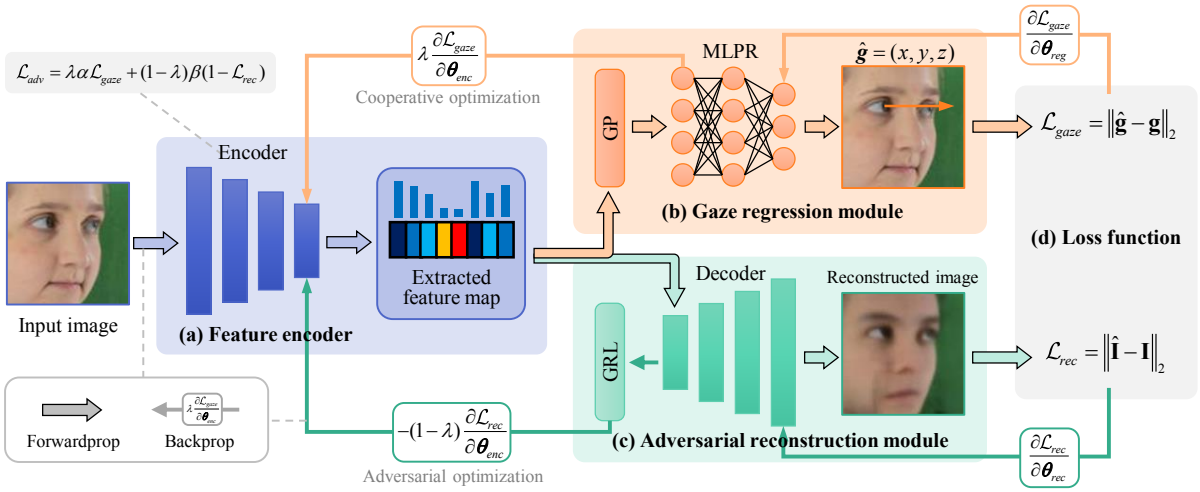


Fig. 2: Framework of generative self-adversarial learning for gaze estimation. (a) Feature encoder. (b) Gaze regression module. GP is the global pooling layer. (c) Adversarial reconstruction module. GRL refers to the gradient reversal layer. (d) Loss function. In the gaze regression task, the gaze regression module performs cooperative optimization with the feature encoder, encouraging to extract more gaze-relevant features.

model is encouraged to learn as fewer unique features as possible, which can be expressed as

$$\min_{E(\boldsymbol{\theta})\in S,\, G\in S} (E(\boldsymbol{\theta}) \cup G)\,, \tag{2}$$

where $\min(\cdot)$ means to minimize a set.

The model appears to be adversarial in achieving these two objectives. The gaze regression task necessitates the model to extract a greater number of gaze features, whereas image reconstruction incentivizes the model to extract fewer features from samples. Through the adversarial learning, combining Eq.1 and Eq.2, the optimization objectives of self-adversarial learning can be summarized as

$$\min_{E(\boldsymbol{\theta})\in S,\, G\in S} ((E(\boldsymbol{\theta}) - G) \cup (G - E(\boldsymbol{\theta})))\,. \tag{3}$$

It effectively measures how well E and G match, guiding the feature encoder to acquire more generalized gaze features and fewer domain-specific ones, thereby enhancing its capacity for domain generalization without requiring any samples from target domains.

B. Framework of GSA-Gaze

We design the GSA-Gaze based on the proposed generative self-adversarial learning mechanism. The framework is shown in Fig. 2, GSA-Gaze comprises the feature encoder, the gaze regression module, the adversarial reconstruction module, and employs the self-adversarial loss for optimization. First, the feature encoder extracts a feature map from the input image, which is subsequently fed into both the gaze regression and generative reconstruction modules. Second, the gaze regression module regresses the gaze vector from the extracted feature map. The accuracy of the gaze regression is directly proportional to the number of features in the feature map that are relevant to the gaze. It engages in cooperative optimization with the feature encoder, thereby incentivizing it to extract more gaze-relevant features. Third, the adversarial reconstruction module reconstructs the input image from the extracted feature map, more unique features of the input image contained in the feature map, the more similar the reconstructed image is to the original image. The adversarial reconstruction module is designed to guide the model in eliminating unique features, thus it is not expected to precisely reconstruct the input image from the extracted feature map. In training, through the gradient reversal layer(GRL), the gradient between the encoder and the decoder is reversed in the back propagation, thus the feature decoder performs adversarial optimization with the feature encoder to encourage extracting fewer domain-specific features.

Through the aforementioned design, under the supervision of both gaze regression and generative reconstruction modules, the feature encoder undergoes self-adversarial learning to acquire general gaze features while eliminating unique ones, enhancing the domain generalization capability of the model.

C. Pyramid-grouped Convolution Encoder

Generative self-adversarial learning mechanism focuses on preserving the general gaze feature and removing other information, in other words, it purifies the extracted features. For GSA-Gaze, the input face image is information-sparse and the gaze patterns are sparsely represented in the input image. The dense channel connections in the standard convolution operation hinder our model from obtaining the purified feature representation. Therefore, in the design of the feature encoder network, we propose the pyramid-grouped convolution (PGC) mechanism to address the channel connection redundancy problem. As Fig. 3 shows, in each pyramid-grouped convolutional block, the density of channel con-

nection changes from dense to sparse. Compared to existing grouped convolution[13], the kernels in the pyramid-grouped convolution possess varying channel-wise receptive fields, which enable the block to focus on the features of every single channel with a longer path, providing a more sparse feature encoding and enhancing the diagonal correlation between channels.

We utilize the PGC blocks to reconstruct ResNet-50[14] as ResPGC-50, which serves as our feature encoder. It extracts the feature map from the input image, expressed as

$$\boldsymbol{Z} = f_{enc}(\boldsymbol{I} \,|\, \boldsymbol{\theta}_{enc}) \,, \tag{4}$$

where $\boldsymbol{Z}$ is the extracted feature map, $f_{enc}(\cdot)$ means the feature encoder, $\boldsymbol{I}$ indicates the input image and $\boldsymbol{\theta}_{enc}$ is the learnable parameters of the encoder.

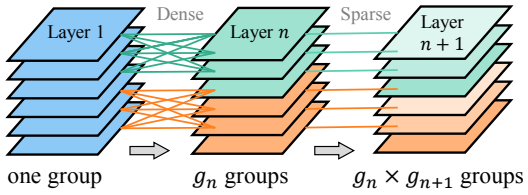


Fig. 3: Pyramid-grouped convolution. In each pyramid block, the density of channel connection changes from dense to sparse.

D. Self-adversarial Loss Function and Optimization

To optimize GSA-Gaze, distinct loss functions are employed for each module within the network. First, for gaze regression module, the mean square error of the gaze vector is used to evaluate the precision of gaze prediction, $\mathcal{L}_{gaze}$ is calculated by

$$\mathcal{L}_{gaze} = \|\hat{\boldsymbol{g}} - \boldsymbol{g}\|_2 \,, \tag{5}$$

where $\hat{\boldsymbol{g}}$ and $\boldsymbol{g}$ are the predicted gaze vector and the ground truth.

Second, for the generative reconstruction module, the reconstructor is encouraged to mine the information in the extracted feature map as much as possible, so the pixel-wise mean square error of the generated image and the input image is used to optimize the reconstructor, $\mathcal{L}_{rec}$ is calculated by

$$\mathcal{L}_{rec} = \|\hat{\boldsymbol{I}} - \boldsymbol{I}\|_2 \,, \tag{6}$$

where $\hat{\boldsymbol{I}}$ and $\boldsymbol{I}$ are the generated image and the input image.

Third, for the feature encoder, as Fig. 2 shows, on the one hand, the extracted feature map is encouraged to support the gaze regression, it is cooperatively optimized with the gaze regression module, so $\mathcal{L}_{gaze}$ is also used for its optimization. On the other hand, the self-adversarial learning makes it expected to unfavorable to precise image reconstruction, it performs the adversarial optimization with the reconstruction module, the adversarial reconstruction loss $1 - \mathcal{L}_{rec}$ is used for its optimization. In summary, the adversarial loss function for the feature encoder $\mathcal{L}_{enc}$ is designed as:

$$\mathcal{L}_{enc} = \mathcal{L}_{adv} = \lambda\alpha\mathcal{L}_{gaze} + (1-\lambda)\beta(1-\mathcal{L}_{rec}) \,, \tag{7}$$

it consists of a positive gaze loss and a reversed reconstruction loss, to satisfy the self-adversarial learning. λ is a hyper parameter used to adjust the bias of the model to the two task, the larger λ is, the model is more biased towards extracting more gaze-relevant features, the smaller λ is, the model is more biased towards eliminating more unique features of training data. λ is set to 0.6 by default. α and β are scale coefficients, which are used to adjust $\mathcal{L}_{gaze}$ and $\mathcal{L}_{rec}$ to an appropriate and uniform scale, the values of them depend on the data preprocessing method. According to the above loss function design, the optimization objectives of the network parameters are as follows:

$$\boldsymbol{\theta}^*_{reg} = \arg\min_{\boldsymbol{\theta}_{reg}} \mathcal{L}_{gaze}(\boldsymbol{\theta}_{enc}, \boldsymbol{\theta}_{reg}), \tag{8}$$

$$\boldsymbol{\theta}^*_{rec} = \arg\min_{\boldsymbol{\theta}_{rec}} \mathcal{L}_{rec}(\boldsymbol{\theta}_{enc}, \boldsymbol{\theta}_{rec}), \tag{9}$$

$$\begin{aligned} \boldsymbol{\theta}^*_{enc} = &\arg\min_{\boldsymbol{\theta}_{enc}} (\lambda\alpha\mathcal{L}_{gaze}(\boldsymbol{\theta}_{enc}, \boldsymbol{\theta}_{reg}) + \\ &(1-\lambda)\beta(1-\mathcal{L}_{rec}(\boldsymbol{\theta}_{enc}, \boldsymbol{\theta}_{rec}))), \end{aligned} \tag{10}$$

where $\boldsymbol{\theta}^*_{reg}$, $\boldsymbol{\theta}^*_{rec}$ and $\boldsymbol{\theta}^*_{enc}$ are the optimized parameters of the gaze regressor, the image reconstructor and the feature encoder.

IV. EXPERIMENTS

A. Experimental Setup

Data preparation: Considering the standardization of data collection, image quality, the number of samples and the diversity of gaze range, we choose Gaze360[11](G) and ETH-XGaze[15](E) as the training sets. MPIIGaze[16](M) and EyeDiap[17](D) are used as testing sets to evaluate the domain generalization ability of models. Using these datasets, we establish four domain generalization tasks for our experiments: G→M, G→D, E→M, E→D. The left side of the arrow is the source domain, and the right side of the arrow is the target domain.

Compared methods: In order to prove the effectiveness of the proposed method, we introduce impactful gaze estimation methods for comparison, including RT-Gene[18], Dilated-Net[5], CA-Net[19]. In addition, to study the domain generalization effects of generative self-adversarial learning, we also used some domain adaption methods for comparison, FSA-Gaze[20], UMA[21], PNP-GA[12].

B. Comparative Results

The angle difference between the estimated gaze vector and the ground truth is used as the evaluation metric. The comparative results are shown in Table I, it proves that existing methods show an obvious performance drop in cross-domain testing. In the without-adaption group, RT-Gene and Dilated-Net even completely fail in four tasks. Our method outperforms the compared methods in all four tasks. In the with-adaption group, 1000 target domain samples are allowed to be used for finetuning, so that we can perform the domain adaption methods. With finetuning, our method show better performance, because samples from the target domain bring features unique to the target domain, enhancing the

TABLE I: Comparative Results on the domain-generalized gaze estimation task across Gaze360(G), ETH-XGaze(E), MPIIGaze(M) and EyeDiap(D).

Adaption	Models	Target domain samples	Gaze angle difference ↓			
			G→M	G→D	E→M	E→D
w/o	RT-Gene[18]	-	22.70°	36.59°	29.84°	31.20°
	Dilated-Net[5]	-	18.45°	23.88°	10.29°	16.74°
	CA-Net[19]	-	12.05°	15.66°	8.94°	8.72°
	GSA-Gaze(ours)	-	**9.83°**	**10.02°**	**7.62°**	**8.14°**
w/	FSA-Gaze[20]	1000	9.03°	10.54°	7.95°	8.22°
	UMA[21]	1000	7.40°	12.65°	7.21°	7.52°
	PNP-GA[12]	1000	**6.25°**	12.90°	6.92°	7.66°
	GSA-Gaze(ours)	1000	6.37°	**8.95°**	**6.45°**	**7.25°**

feature representation of target domain data. In addition, in the G→M task, PNP-GA has the best performance, in other three tasks, our method outperforms other methods in with-adaption group. The results support that our methods have batter generalized gaze estimation capability in both with-adaption and without-adaption cases.

C. Ablation Analysis

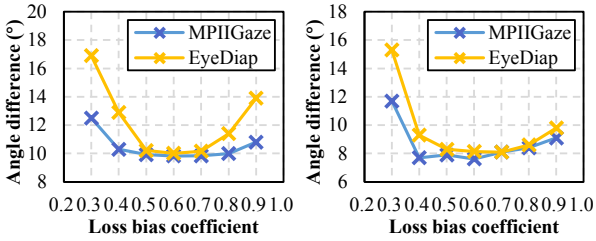
To verify the effectiveness of the proposed generative self-adversarial learning, we designed three groups of ablation experiments. The generative self-adversarial learning(GSAL) mechanism is portable, by adding the reconstruction module and self-adversarial loss function, we can easily plug the mechanism into other methods. In group 1, we plug the the GSAL to Dilated-Net. In group 2, we plug the GSAL to a gaze regression model based on ResNet-50[14]. In group 3, we remove the adversarial reconstruction module and the self-adversarial loss in our methods, and use it as the baseline. The ablation analysis results are shown in Table II. In all 3 groups and all 4 tasks, the addition of the GSAL mechanism makes the baseline model show better performance, which proves that the proposed GSAL mechanism can also effectively enhance the domain generalization capability of other gaze estimation methods. Additionally, GSA-Gaze shows better performance than ResNet-50 + GSAL, the backbone of the feature encoder in GSA-Gaze is ResPGC-50, so the findings also provide evidence that pyramid-grouped convolution plays a role in mitigating cross-domain feature over-fitting, thereby enhancing the performance of generalized gaze estimation.

D. Loss Bias Coefficient Analysis

When optimizing the parameters of the feature encoder in GSA-Gaze, the loss bias coefficient λ is utilized to adjust the bias of the model to gaze regression task and image reconstruction task, as Eq. (7) shows. To study the effect of λ value to model performance on generalized gaze estimation, we train the model with different λ values, and perform cross-domain testing. The results are shown in Fig. 4, as the value of λ increases, the average angle difference first decreases and then increases. When $\lambda = 0.6$, the model exhibits the best performance, when λ is too small or too large, the model shows a performance drop. The results are consistent with the anticipated outcome, in extreme conditions, if $\lambda = 1$, the reconstruction loss will be ignored during training and adversarial learning will not work, the model is equivalent to the baseline, if $\lambda = 0$, the gaze regression loss will be disregarded, the model will be unable to perform accurate gaze estimation. Therefore, in the above experiments in this paper, we set λ to 0.6 to obtain the best performance.

TABLE II: Ablation analysis results on the domain-generalized gaze estimation task across Gaze360(G), ETH-XGaze(E), MPIIGaze(M) and EyeDiap(D).

Methods	Gaze angle difference ↓			
	G→M	G→D	E→M	E→D
Dilated-Net	18.45°	23.88°	10.29°	16.74°
Dilated-Net + GSAL	10.66°	16.48°	10.01°	9.93°
ResNet-50	11.79°	16.79°	10.67°	10.02°
ResNet-50 + GSAL	10.57°	12.69°	7.80°	8.73°
ResPGC-50(ours)	11.25°	14.27°	9.66°	10.39°
GSA-Gaze(ours)	**9.83°**	**10.02°**	**7.62°**	**8.14°**



(a) Training on Gaze360 (b) Training on ETH-XGaze

Fig. 4: Effect of λ value to model performance on generalized gaze estimation. (a) Results of training on Gaze360 and testing on MPIIGaze and EyeDiap. (b) Results of training on ETH-XGaze and testing on MPIIGaze and EyeDiap.

E. Visualized Reconstruction Results

The fundamental principle of generative self-adversarial learning involves integrating an adversarial image reconstruction task to compel the model to utilize fewer features for gaze estimation, thereby reducing feature redundancy and emphasizing only on general gaze features, thus preventing cross-domain feature over-fitting. To gain a better understanding of the GSAL mechanism, we employ visualizations to the images reconstructed from the extracted feature maps, as Fig. 5 shows. The top row displays the original images, while the bottom row shows the reconstructed images. The reconstructed images exhibit minimal information regarding personal characteristics, while retaining the features pertaining to gaze. The decoder tends to reconstruct an image that resembles an average face, which is very different from the original image, however, the structural information of the face and eyes related to the gaze is retained. Specifically, the reconstructed image in the third column does not retain the illumination characteristics of the original image. In the fourth column, the glasses in the original image are eliminated from the extracted features. In the fifth column, the bread in the original face is removed. These findings demonstrate the efficacy of GSAL in eliminating domain-specific features while preserving general gaze characteristics.

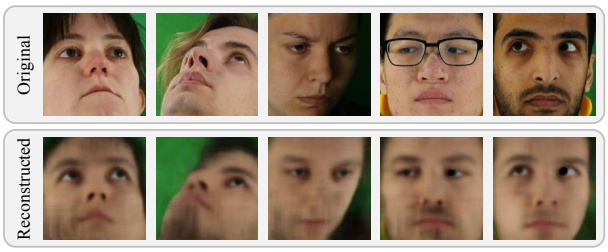


Fig. 5: Visualized reconstruction results. The top row shows the original images, the bottom row shows the reconstructed images.

V. CONCLUSION

In this paper, we propose a novel approach for domain generalization in gaze estimation based on generative self-adversarial learning. The Experiments results show that the proposed generative self-adversarial learning method can effectively guide the model to learn general gaze features while eliminating the domain-specific features, significantly enhancing the domain generalization capability of gaze estimation model, which helps advance the application of gaze estimation in intelligent transportation systems.

REFERENCES

[1] J. Wu, Q. Kong, K. Yang, Y. Liu, D. Cao, and Z. Li, "Research on the steering torque control for intelligent vehicles co-driving with the penalty factor of human–machine intervention," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 53, no. 1, pp. 59–70, 2022.

[2] Y. Cheng, H. Wang, Y. Bao, and F. Lu, "Appearance-based gaze estimation with deep learning: A review and benchmark," *arXiv preprint arXiv:2104.12668*, 2021.

[3] P. Pathirana, S. Senarath, D. Meedeniya, and S. Jayarathna, "Eye gaze estimation: A survey on deep learning-based approaches," *Expert Systems with Applications*, vol. 199, p. 116894, 2022.

[4] A. Farahani, S. Voghoei, K. Rasheed, and H. R. Arabnia, "A brief review of domain adaptation," *Advances in Data Science and Information Engineering: Proceedings from ICDATA 2020 and IKE 2020*, pp. 877–894, 2021.

[5] Z. Chen and B. E. Shi, "Appearance-based gaze estimation using dilated-convolutions," in *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part VI*. Springer, 2019, pp. 309–324.

[6] Y. Cheng, X. Zhang, F. Lu, and Y. Sato, "Gaze estimation by exploring two-eye asymmetry," *IEEE Transactions on Image Processing*, vol. 29, pp. 5259–5272, 2020.

[7] Y. Bao, Y. Cheng, Y. Liu, and F. Lu, "Adaptive feature fusion network for gaze tracking in mobile tablets," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 9936–9943.

[8] Y. Zheng, S. Park, X. Zhang, S. De Mello, and O. Hilliges, "Self-learning transformations for improving gaze and head redirection," *Advances in Neural Information Processing Systems*, vol. 33, pp. 13 127–13 138, 2020.

[9] Y. Ganin and V. Lempitsky, "Unsupervised domain adaptation by backpropagation," in *International conference on machine learning*. PMLR, 2015, pp. 1180–1189.

[10] K. Wang, R. Zhao, H. Su, and Q. Ji, "Generalizing eye tracking with bayesian adversarial learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 11 907–11 916.

[11] P. Kellnhofer, A. Recasens, S. Stent, W. Matusik, and A. Torralba, "Gaze360: Physically unconstrained gaze estimation in the wild," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 6912–6921.

[12] Y. Liu, R. Liu, H. Wang, and F. Lu, "Generalizing gaze estimation with outlier-guided collaborative adaptation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 3835–3844.

[13] S. Xie, R. Girshick, P. Dollár, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1492–1500.

[14] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[15] X. Zhang, S. Park, T. Beeler, D. Bradley, S. Tang, and O. Hilliges, "Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*. Springer, 2020, pp. 365–381.

[16] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Mpiigaze: Real-world dataset and deep appearance-based gaze estimation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 1, pp. 162–175, 2017.

[17] K. A. Funes Mora, F. Monay, and J.-M. Odobez, "Eyediap: A database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras," in *Proceedings of the symposium on eye tracking research and applications*, 2014, pp. 255–258.

[18] T. Fischer, H. J. Chang, and Y. Demiris, "Rt-gene: Real-time eye gaze estimation in natural environments," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 334–352.

[19] Y. Cheng, S. Huang, F. Wang, C. Qian, and F. Lu, "A coarse-to-fine adaptive network for appearance-based gaze estimation," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 10 623–10 630.

[20] S. Park, S. D. Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz, "Few-shot adaptive gaze estimation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 9368–9377.

[21] M. Cai, F. Lu, and Y. Sato, "Generalizing hand segmentation in egocentric videos with uncertainty-guided model adaptation," in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, 2020, pp. 14 392–14 401.