Supplementary Material

A. Proofs

A.1 Proof for Proposition 1

Proposition 1. *We derive the new contrastive loss function for regression task as*

$$-\log \frac{\sum_k \mathbb{S}[p(g_i); p(g_k)] \cdot f_k(y_k, x)}{\sum_j f_j(y_j, x)} \tag{S-1}$$

where $f_i(y_i, x)$ is the density ratio.

Proof. Following InfoNCE [10], we define $\frac{p(y|x)}{p(y)}$ as the density ratio, where $p(y|x)$ is the predict distribution that we want, while $p(y)$ is the noise distribution used for contrast. Considering N as the batch size, the probability of finding the positive sample $p(y_i|x)$ as:

$$\frac{p\left(y_i|x\right) \prod_{l \neq i} p\left(y_l\right)}{\sum_{j=1}^{N} p\left(y_j|x\right) \prod_{l \neq j} p\left(y_l\right)} = \frac{\frac{p(y_i|x)}{p(y_i)}}{\sum_{j=1}^{N} \frac{p(y_j|x)}{p(y_j)}} \tag{S-2}$$

According to $f_i(y_i, x) \propto \frac{p(y_i|x)}{p(y_i)}$ [10], we have the predict distribution as

$$p(y_i|x) = \frac{f_i(y_i, x)}{\sum_j f_j(y_j, x)} \tag{S-3}$$

For regression model, different from classification tasks, the relationship between labels reveal the relationship between the features. Then we can make a assumption that the ratio between predict distribution $p(y_i|x)$ and $p(y_k|x)$ is proportional to the similarity between label distribution $p(g_i)$ and $p(g_j)$. Then we have

$$\frac{p(y_i|x)}{p(y_k|x)} = \frac{f_i(y_i, x)}{f_k(y_k, x)} = C \cdot \mathbb{S}[p(g_i); p(g_k)] \tag{S-4}$$

In other words, $f_i(y_i, x) = C \cdot \mathbb{S}[p(g_i); p(g_k)] \cdot f_k(y_k, x)$. Considering other samples, we have N similar expressions. We take the sum we get the predict distribution as:

$$p(y_i|x) = \frac{\frac{C}{N} \sum_k \mathbb{S}[p(g_i); p(g_k)] \cdot f_k(y_k, x)}{\sum_j f_j(y_j, x)} \tag{S-5}$$

Following the MLE loss function, we have the derived new loss function as Contrastive Regression loss (CR loss):

$$\begin{aligned} \mathcal{L} &= -\log \frac{\frac{C}{N} \sum_k \mathbb{S}[p(g_i); p(g_k)] \cdot f_k(y_k, x)}{\sum_j f_j(y_j, x)} \\ &= -\log \frac{\sum_k \mathbb{S}[p(g_i); p(g_k)] \cdot f_k(y_k, x)}{\sum_j f_j(y_j, x)} - \log \frac{C}{N} \end{aligned} \tag{S-6}$$

As the last term in Eq. S-6 is a constant and can be omitted:

$$-\log \frac{\sum_k \mathbb{S}[p(g_i); p(g_k)] \cdot f_k(y_k, x)}{\sum_j f_j(y_j, x)} \tag{S-7}$$

□

A.2 Proof for Proposition 2

Proposition 2. *The two forms of loss function $L_1 = -\log \frac{\sum_k \mathbb{S}_{i,k} \cdot f_k(y_k, x)}{\sum_j f_j(y_j, x)}$ and $L_2 = -\log \frac{\sum_k \sigma(\mathbb{S}_{i,k}) \cdot f_k(y_k, x)}{\sum_j |\mathbb{S}_{i,k}| \cdot f_j(y_j, x)}$ have the same effect, i.e., they pull features with closer gaze directions closer together while pushing features with farther gaze directions farther apart.*

Proof. Considering that the $\log(\cdot)$ function is monotonically increasing, we consider the gradient for the inner function. For $I_1 = \frac{\sum_k \mathbb{S}_{i,k} \cdot f_k(y_k, x)}{\sum_j f_j(y_j, x)}$, we have the gradient for $f_m(y_m, x)$ is

$$\begin{aligned} \frac{\partial I_1}{\partial f_m} &= \frac{\mathbb{S}_{i,m} \cdot \sum_j f_j(y_j, x) - \sum_k \mathbb{S}_{i,k} \cdot f_k(y_k, x)}{[\sum_j f_j(y_j, x)]^2} \\ &= \frac{\sum_k (\mathbb{S}_{i,m} - \mathbb{S}_{i,k}) \cdot f_k(y_k, x)}{[\sum_j f_j(y_j, x)]^2} \end{aligned} \tag{S-8}$$

Considering that $f_k(y_k, x) \geq 0$ always holds, then $\mathbb{S}_{i,m} - \mathbb{S}_{i,k}$ determines the direction of the gradient. If $\mathbb{S}_{i,m}$ is larger than $\mathbb{S}_{i,k}$, then we need to enlarge $f_m(y_m, x)$ to maximize the inner function I_1, equivalent to minimize the L_1 loss, while $\mathbb{S}_{i,m}$ is smaller than $\mathbb{S}_{i,k}$, we need to reduce $f_m(y_m, x)$. This indicates that we will pull features with closer labels closer together while push features with further labels further apart.

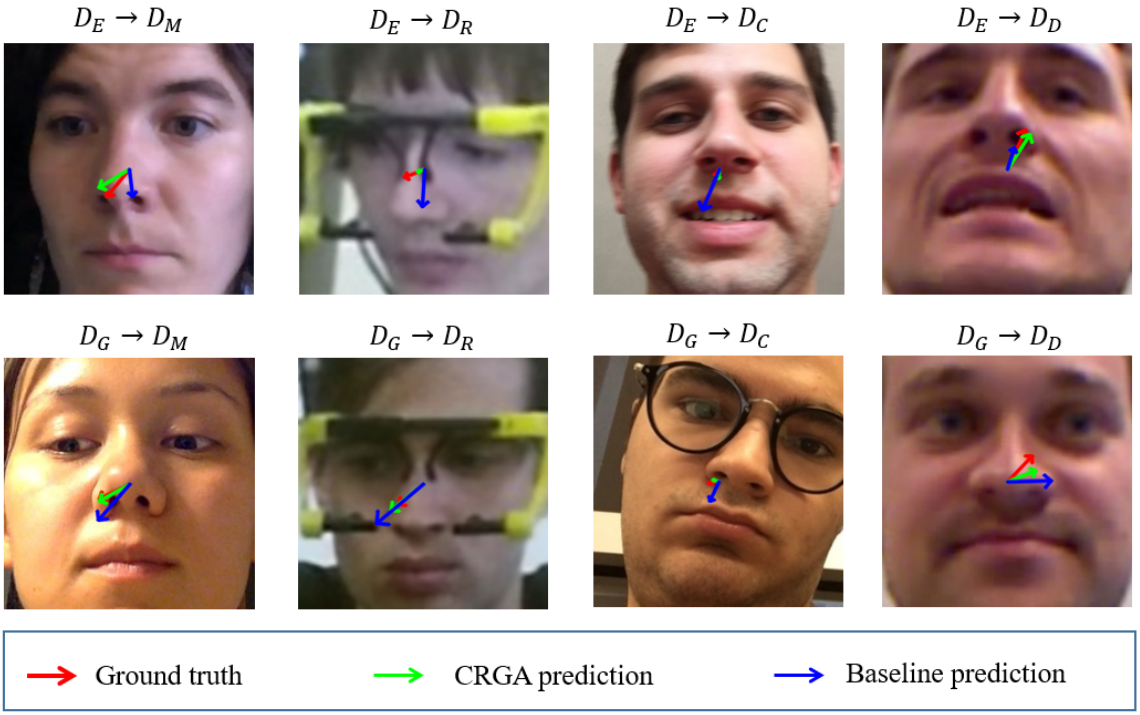


Figure S-1. Visual examples of estimated 3d gaze result. Arrows represent the projection of the 3d unit gaze vector on the image plane. Red arrow represent the ground-truth gaze vector, green arrow represent the predictions of CRGA and blue arrow represent the predictions of baseline model.

For $I_2 = \frac{\sum_k \sigma(\mathbb{S}_{i,k}) \cdot f_k(y_k,x)}{\sum_j |\mathbb{S}_{i,j}| \cdot f_j(y_j,x)}$, where σ is the relu function. We have the gradient for $f_m(y_m, x)$ is

$$\begin{aligned}\frac{\partial I_2}{\partial f_m} &= \frac{\sigma(\mathbb{S}_{i,m}) \sum_j |\mathbb{S}_{i,j}| f_j - \sum_k \sigma(\mathbb{S}_{i,k}) f_k |\mathbb{S}_{i,m}|}{[\sum_j |\mathbb{S}_{i,j}| \cdot f_j]^2} \\ &= \frac{\sum_j [\sigma(\mathbb{S}_{i,m}) \cdot |\mathbb{S}_{i,j}| - \sigma(\mathbb{S}_{i,j}) \cdot |\mathbb{S}_{i,m}|] \cdot f_j}{[\sum_j |\mathbb{S}_{i,j}| \cdot f_j]^2}\end{aligned} \tag{S-9}$$

Then $\sigma(\mathbb{S}_{i,m}) \cdot |\mathbb{S}_{i,j}| - \sigma(\mathbb{S}_{i,j}) \cdot |\mathbb{S}_{i,m}|$ determines the direction of the gradient. As we elaborated below in Sec.A.3, we take $\mathbb{S}_{i,j} \geq 0$ as the close pairs (the assumed constant variance 0.07 is small and ensures few positive pairs). Then we consider the scene where $\mathbb{S}_{i,j}$ and $\mathbb{S}_{i,m}$ with opposite symbols. Then if $\mathbb{S}_{i,m} \geq 0$, then $\sigma(\mathbb{S}_{i,m}) \cdot |\mathbb{S}_{i,j}| - \sigma(\mathbb{S}_{i,j}) \cdot |\mathbb{S}_{i,m}| \geq 0$ and will $\nearrow$ as $\mathbb{S}_{i,m} \nearrow$ While $|\mathbb{S}_{i,j}| - \sigma(\mathbb{S}_{i,j}) \cdot |\mathbb{S}_{i,m}| \leq 0$ and will $\searrow$ as $\mathbb{S}_{i,m} \nearrow$ when $\mathbb{S}_{i,m} \leq 0$. This also indicates that we will pull features with closer labels closer together while push features with further labels further apart. □

A.3 Similarity Function

As we can see from the gaze direction distribution in Fig.3, the gaze direction is mainly concentrated in front of face and most of the datasets are concentrated with less difference in gaze directions. This means we need to distinguish subtle differences between gaze directions.

If we choose cosine similarity as the similarity function $\mathbb{S}_{i,j}$ between label distribution $p(g_i)$ and $p(g_j)$, the low gradient near zero will exacerbate the difficulty in distinguishing gaze directions. Thus we propose to use the negative log KL divergence to model the difference between label distribution.

Following [8], We assume that the density follows the Laplace distribution with a constant variance as $g \sim La(g; \mu, \delta)$. For this constant variance, we take $\delta = 0.07$. Here, $0.07 \approx 4°/180° \cdot \pi$, where $4°$ is our assumed standard estimation for personal error between visual axis and optical axis. Then the Kl divergence between $p(g_i) = La(\mu_i, \delta_i)$ and $p(g_j) = La(\mu_j, \delta_j)$ is

$$\mathcal{D}_{KL}[p(g_i)||p(g_j)] = |\mu_i - \mu_j| \tag{S-10}$$

Because kl divergence tends to zero when the two distributions are similar, which is the opposite of our goal, we take negative kl. To further encourage the difference near zero, we take the log function and get

$$\mathbb{S}_{i,j} = -\log |\mu_i - \mu_j| \tag{S-11}$$

Then, when $|\mu_i - \mu_j| \leq 1$, we will pull together two gaze features. However, this means gaze direction error as $1/pi \cdot 180 \approx 57°$. This is not reasonable, thus we introduce the comparison with our assumed assumed standard estimation for personal error $4°$. Considering that the gaze direction

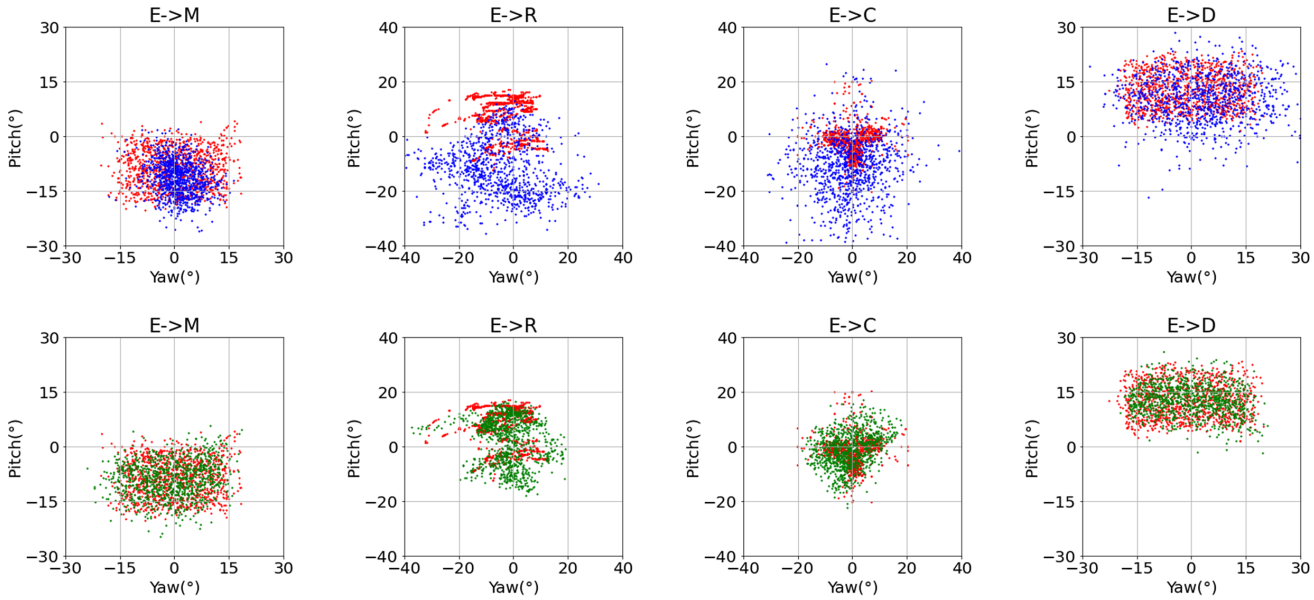


Figure S-2. Scatter plot of predictions and ground-truth labels before and after our CRGA on source domain $\mathcal{D}_E$. Scatter points in blue represent baseline predictions, red ones represent ground-truth labels, and green ones represent predictions of our CRGA.

is mainly concentrated in front of the face and the gradient near zero of cosine similarity is too small, we derive a -log KL function as the similarity:

$$\mathbb{S}_{i,j} = -\log \frac{|\mu_i - \mu_j|}{0.07} = \log \frac{0.07}{|\mu_i - \mu_j|} \tag{S-12}$$

Here, μ is the gaze label obtained from the collection.

B. Datasets

ETH-XGAZE [11] is collected with 18 digital SLR cameras from 110 participants in laboratory environments, which contains large variations in head poses, gaze direction, personal feature and illumination condition. It provides 80 subjects ($i.e.$,756,540 images) as the training set.

Gaze360 [6] is collected in both indoor and outdoor environments, which contains labelled 3D gaze of 238 subjects with a wide-range head pose and gaze direction. Following [2, 9], we remove images without subjects' faces but use the remaining 84,902 images as the training set (different from [2, 9], 16,031 test images are not used as the training set for more fair training).

MPIIGaze [12] is collected from 15 subjects in real-world environments. According to the standard evaluation protocol, which selects 3000 images from each subject to form an evaluation set, we adopt the evaluation set directly.

RT-GENE [3] is collected in natural environment with large camera-to-subject distances, which contains high variations in head poses and gaze as well. Following [2, 9], 108,965 images are employed as the evaluation set.

GazeCapture [7] contains over 2.5 M frames collected from 1450 people, which is collected with mobile phone and tablets. Following [2,9], 179,496 images from 150 subjects are employed as the evaluation set.

EyeDiap [4] contains video clips from 16 subjects and screen targets or 3D floating balls are taken as gaze target. Following [2,9], 16,674 images from 14 subjects under screen target sessions are employed as the evaluation set.

C. Training Details

Training details. We perform our experiments on Tesla-V100 GPU. The resolution for input images in all the experiments is set as 224×224, which follows the convention of [6,11], while different from [9], which employs 448×448 as the resolution of input for training on Gaze360. We take ResNet-50 [5] as the backbone to extract features for all experiments if without extra annotation, a 2-layers MLP as CR predictor to generate 128-dim CR feature vectors, and an FC layer to regress a 2-dim gaze vector for pitch and yaw angles respectively. For domain generalization task CDG on source domain $\mathcal{D}_E$, we follow [11], set the batch size as 128, use the Adam optimizer with a learning rate of 5×10^{-4} and train for 25 epochs using a decay factor 0.1 every 10 epochs. For CDG on source domain $\mathcal{D}_G$, we follow [6], set the batch size as 128, use the Adam optimizer with a learning rate of 4×10^{-4} and train for 100 epochs. For domain adaptation tasks, the hyperparameter setting keeps the same as that in domain generalization task CDG in source domain $\mathcal{D}_E$.

Data augmentations. We employ a data augmentation family with a random color field and greyscale.

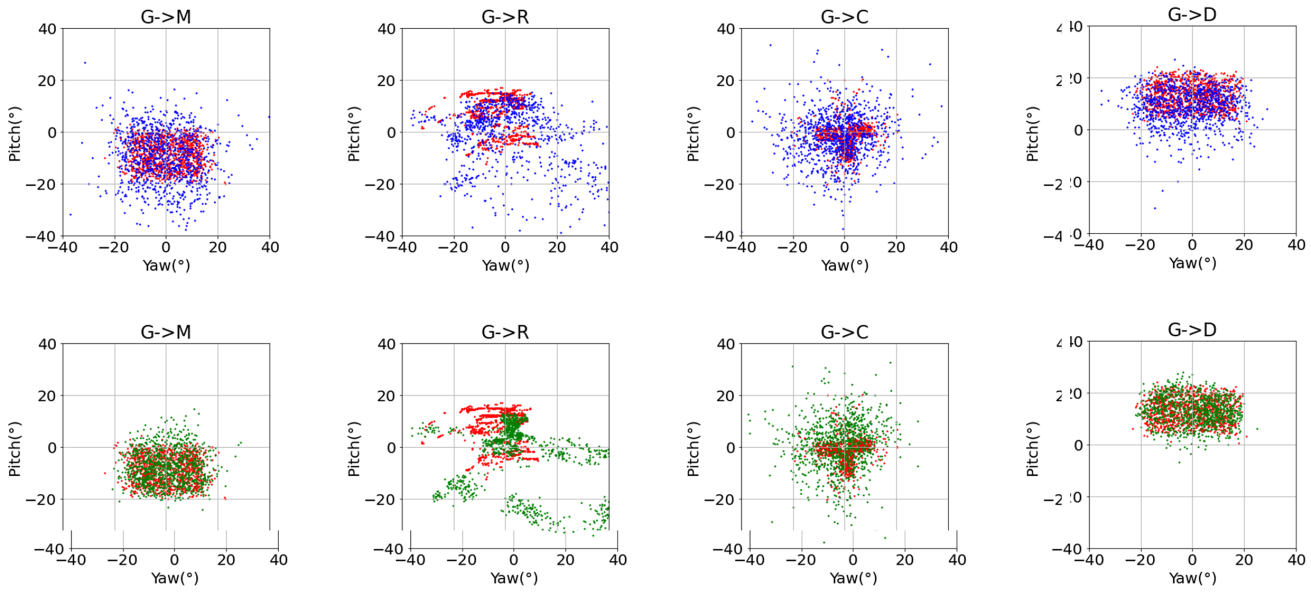


Figure S-3. Scatter plot of predictions and ground-truth labels before and after our CRGA on source domain $\mathcal{D}_G$. Scatter points in blue represent baseline predictions, red ones represent ground-truth labels, and green ones represent predictions of our CRGA.

D. Additional Experiments

D.1 Ablation study on prior λ.

Illustrated in Tab. S-1, high λ will mislead the model to pull feature with far gaze together, while small λ will push feature with close gaze apart. Thus we choose $\lambda = 0.07$ and achieve an impressive performance.

CDG	$\mathcal{D}_G \rightarrow \mathcal{D}_M$	$\mathcal{D}_G \rightarrow \mathcal{D}_R$	$\mathcal{D}_G \rightarrow \mathcal{D}_C$	$\mathcal{D}_G \rightarrow \mathcal{D}_D$
$\lambda = 0.03$	7.55	**20.75**	9.29	7.91
$\lambda = 0.07$	**7.03**	**20.79**	**8.28**	**7.27**
$\lambda = 0.15$	8.05	25.39	9.94	7.91

Table S-1. Ablation study on different λ.

D.2 Comprehension of CRGA

To understand the effectiveness of our proposed CRGA intuitively, we visualize the gaze prediction on eight gaze adaptation task on Fig. S-1. The 3d gaze direction is represented by the projection of 3d unit vector on the image plane as arrow. We compare the gaze ground truth, gaze prediction by CRGA and gaze prediction by baseline model in the figure with red, green and red arrows respectively. The visualization show that the gaze prediction by CRGA is more closer to ground truth and CRGA could alleviate the domain adaptation problem of gaze estimation. We also show the distributions of predictions before and after our CRGA on eight tasks in Fig. S-2 and Fig. S-3. Our method significantly reduces the degree of outlier, the prediction distribution of our CRGA are much closer to the distribution of ground-truth labels.

D.3 Extension Experiments on different backbones

We perform two sets of experiments on the domain adaptation task $\mathcal{D}_G \rightarrow \mathcal{D}_M$ using ResNet-50 as the backbone and $\mathcal{D}_E \rightarrow \mathcal{D}_M$ using ResNet-18 as the backbone respectively. In detail, for each set of experiments, we conduct two pipelines for comparison, one in which we perform our CRGA for different iterations I, the other in which we perform self-training with different iterations I on the baseline model without our derived CSA loss. This indicates that our approach could be suitable and effective for different backbones.

D.4 Extension Experiments on Feature Visualization

We conduct further experiments on feature visualization to exhibits the effectiveness of our proposed CR loss. We perform our experiments for CRGA tasks on $\mathcal{D}_E \rightarrow \mathcal{D}_M$, $\mathcal{D}_E \rightarrow \mathcal{D}_C$, $\mathcal{D}_E \rightarrow \mathcal{D}_D$, $\mathcal{D}_G \rightarrow \mathcal{D}_M$, $\mathcal{D}_G \rightarrow \mathcal{D}_C$, $\mathcal{D}_G \rightarrow \mathcal{D}_D$. Here we do not perform experiments on $\mathcal{D}_R$ for its poor performance compared with other target domains (even though we can outperforms the state-of-art performance). The results are presented in Fig. S-4. All the figures shows a clear colour gradient that features with closer labels are pull closer together.

D.5 Extension Experiments on Decoupling Gaze and Head Pose

We conduct further experiments to decouple the gaze estimation and head pose estimation. Two specific full connect modules are employed to regress gaze and head pose

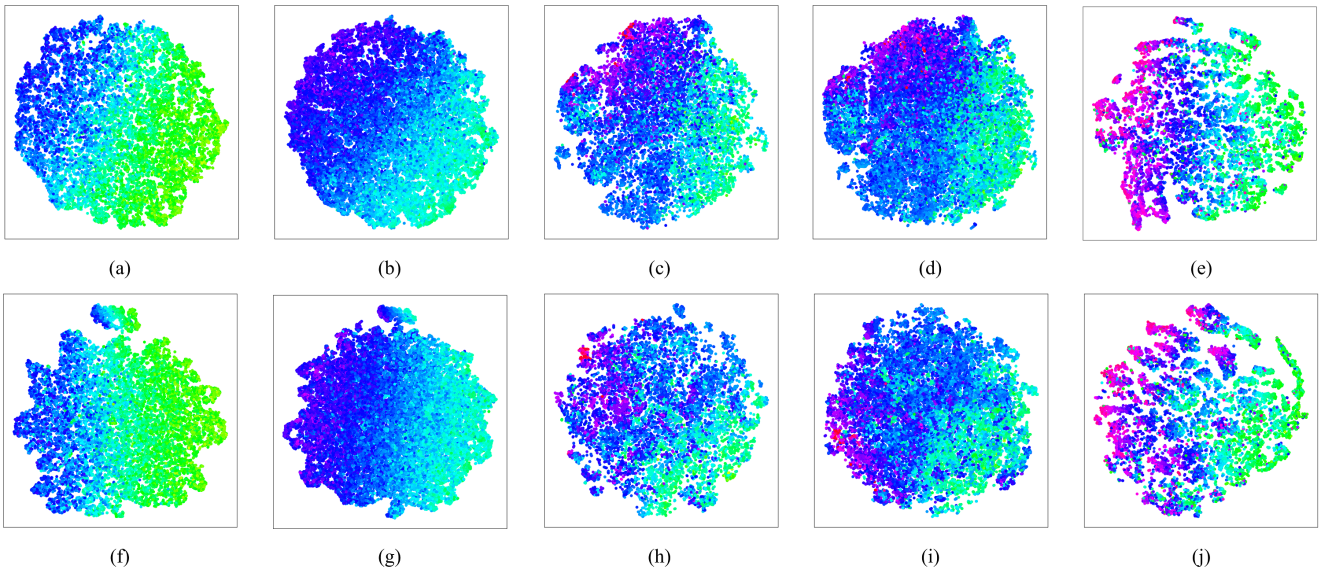
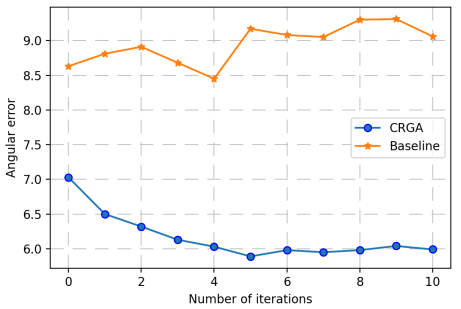
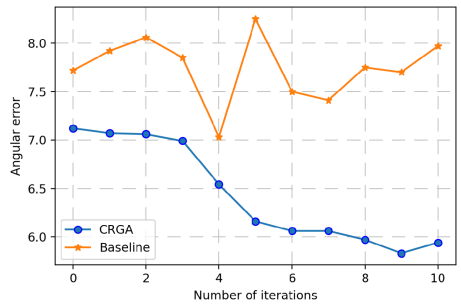


Figure S-4. Illustration of the feature distribution, different colors indicate different gaze directions(best viewed in color). In the first row, we conduct experiments from source domain $\mathcal{D}_E$. (a) is the visualization of 20,000 feature points selected from the datasets on $\mathcal{D}_E \rightarrow \mathcal{D}_M$, (b) is the visualization of all feature points (45,000) from the datasets on $\mathcal{D}_E \rightarrow \mathcal{D}_M$, (c) is the visualization of 20,000 feature points selected from the datasets on $\mathcal{D}_E \rightarrow \mathcal{D}_C$, (d) is the visualization of 50,000 feature points selected from the datasets on $\mathcal{D}_E \rightarrow \mathcal{D}_C$, (e) is the visualization of all feature points (16,674) from the datasets on $\mathcal{D}_E \rightarrow \mathcal{D}_D$. In the second row, we conduct experiments from source domain $\mathcal{D}_G$. (f) is the visualization of 20,000 feature points selected from the datasets on $\mathcal{D}_G \rightarrow \mathcal{D}_M$, (g) is the visualization of all feature points (45,000) from the datasets on $\mathcal{D}_G \rightarrow \mathcal{D}_M$, (h) is the visualization of 20,000 feature points selected from the datasets on $\mathcal{D}_G \rightarrow \mathcal{D}_C$, (i) is the visualization of 50,000 feature points selected from the datasets on $\mathcal{D}_G \rightarrow \mathcal{D}_C$, (j) is the visualization of all feature points (16,674) from the datasets on $\mathcal{D}_G \rightarrow \mathcal{D}_D$.



(a) $\mathcal{D}_G \rightarrow \mathcal{D}_M$ using ResNet50



(b) $\mathcal{D}_E \rightarrow \mathcal{D}_M$ using ResNet18

Figure S-5. Ablation study on iterations of self-training. Angular gaze error (°) is used as the evaluation metric.

Method	$\mathcal{D}_E \rightarrow \mathcal{D}_M$	$\mathcal{D}_E \rightarrow \mathcal{D}_R$	$\mathcal{D}_E \rightarrow \mathcal{D}_C$	$\mathcal{D}_E \rightarrow \mathcal{D}_D$
Baseline	9.19	18.23	13.43	8.62
CDG	**6.73**	**16.45**	9.23	7.85
Decouple	7.36	19.65	**9.04**	**7.18**

Table S-2. Ablation study on decoupling the gaze and head pose. Angular gaze error (°) is used as evaluation metric. Here, lower error rate stands for better performance.

respectively. We choose $\mathcal{D}_E$ as the source domain and verify the domain generation performance. The results is shown in Fig S-2. Decoupling the gaze and head pose exhibits slight improvement generalized on $\mathcal{D}_C$ and $\mathcal{D}_D$ while degrading generalized on $\mathcal{D}_M$ and $\mathcal{D}_R$. This indicates that decoupling gaze and head pose directly provides little help.

D.6 Extension Experiments on our motivation

We test unsupervised contrastive learning method DINO [1], but it only gets a 45 ° gaze error on $\mathcal{D}_E$. We visualize the attention map from 5 heads of the VIT trained with supervised and unsupervised manner (DINO) in Fig S-6, and almost no eye attention is captured by DINO. Besides, the supervised contrastive classification learning (SupCon)

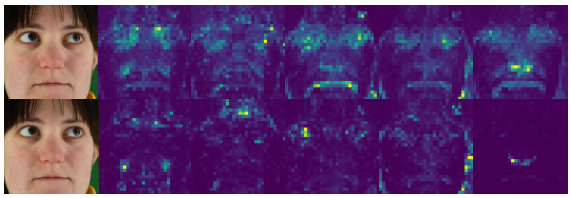


Figure S-6. Attention maps from five heads of vision transformers. The top is supervised, while the bottom is unsupervised (DINO).

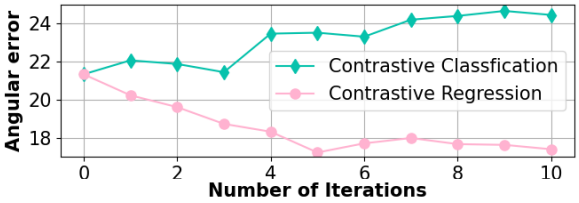


Figure S-7. Comparison of SupCon loss and CR loss on head pose regression domain adaptation on $\mathcal{D}_E \rightarrow \mathcal{D}_R$.

failed in regression tasks. Other than gaze estimation, we further conduct self-training experiments on head pose domain adaptation $\mathcal{D}_E \rightarrow \mathcal{D}_R$ in Fig. S-7 to prove that our motivation is not limited to gaze.

E. Limitation and Future Work

Although CRGA facilitates gaze domain adaptation, it still has limitations. Such as several iterations for self-training, which could be further explored with the EMA teacher. The outliers in the visualization of our feature distribution suggest that we can make further improvements on aligning the features in future research. Further exploration of the contrastive regression learning on the source domain in a totally unsupervised manner (not the domain adaptation) is left for future research.

References

[1] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV 2021 - International Conference on Computer Vision*, 2021. 5

[2] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *arXiv preprint arXiv:2104.12668*, 2021. 3

[3] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 334–352, 2018. 3

[4] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc 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*, pages 255–258, 2014. 3

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3

[6] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6912–6921, 2019. 3

[7] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016. 3

[8] Jiefeng Li, Siyuan Bian, Ailing Zeng, Can Wang, Bo Pang, Wentao Liu, and Cewu Lu. Human pose regression with residual log-likelihood estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11025–11034, 2021. 2

[9] Yunfei Liu, Ruicong Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with outlier-guided collaborative adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3835–3844, 2021. 3

[10] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 1

[11] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *European Conference on Computer Vision*, pages 365–381. Springer, 2020. 3

[12] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE transactions on pattern analysis and machine intelligence*, 41(1):162–175, 2017. 3