Bootstrap AutoEncoders With Contrastive Paradigm for Self-supervised Gaze Estimation

Yaoming Wang [1] **Jin Li** [1] **Wenrui Dai** [1] **Bowen Shi** [1] **Xiaopeng Zhang** [2] **Chenglin Li** [1] **Hongkai Xiong** [1]
{wang_yaoming, deserve_lj, daiwenrui, sjtu_shibowen,
lcl1985, xionghongkai}@sjtu.edu.cn; zxphistory@gmail.com

Abstract

Existing self-supervised methods for gaze estimation using the dominant streams of contrastive and generative approaches are restricted to eye images and could fail in general full-face settings. In this paper, we reveal that contrastive methods are ineffective in data augmentation for self-supervised full-face gaze estimation, while generative methods are prone to trivial solutions due to the absence of explicit regularization on semantic representations. To address this challenge, we propose a novel approach called **B**ootstrap auto-**e**ncoders with **C**ontrastive p**a**radigm (**BeCa**), which combines the strengths of both generative and contrastive methods. Specifically, we revisit the Auto-Encoder used in generative approaches and incorporate the contrastive paradigm to introduce explicit regularization on gaze representation. Furthermore, we design the InfoMSE loss as an alternative to the vanilla MSE loss for Auto-Encoder to mitigate the inconsistency between reconstruction and representation learning. Experimental results demonstrate that the proposed approaches outperform state-of-the-art unsupervised gaze approaches on extensive datasets (including wild scenes) under both within-dataset and cross-dataset protocols.

1. Introduction

Gaze estimation is a vital visual task that predicts the direction of a person's gaze, a crucial non-verbal cue in various applications, including human-computer interaction (Hutchinson et al., 1989; Kim & Ramakrishna, 1999; Surakka et al., 2004; Lei et al., 2023), autonomous vehicles (Martin et al., 2018; Gerber et al., 2020; Pal et al., 2020; Fletcher & Zelinsky, 2009), and augmented and virtual reality (Patney et al., 2016; Burova et al., 2020; Konrad et al., 2020; Shi et al., 2020; Liu & Qin, 2022; Chen et al., 2022a). With the advent of deep learning, appearance-based approaches (Cheng et al., 2021; Zhang et al., 2015; Lu et al., 2014; Liu et al., 2021; Cheng & Lu, 2023; Xu et al., 2023b), regressing gaze direction with neural networks in a supervised manner, have become dominant in gaze estimation.

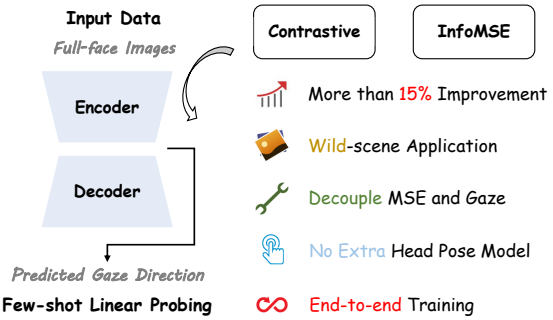


Figure 1: **Characteristics of BeCa and BeCa-InfoMSE.** We introduce contrastive regularization and Informative Mean Square Error to self-supervised gaze estimation with more than 15% performance improvement and well accommodation to wild datasets like Gaze360. BeCa is free from additional head pose models and can be trained in an end-to-end manner. BeCa-InfoMSE further enhances estimation performance by decoupling the learned gaze representation from the reconstruction quality.

Early appearance-based approaches focused on utilizing eye images by assuming that the eye region contains the most abundant gaze information. (Zhang et al., 2017b; Kellnhofer et al., 2019; Zhang et al., 2020; Murthy & Biswas, 2021) demonstrated that the facial region also contains valuable gaze information and can be leveraged to enhance gaze estimation. However, these appearance-based approaches require large amounts of labeled data and are resource-intensive. To address this limitation, self-supervised gaze estimation has emerged as a promising alternative.

[1]School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai, China [2]Huawei Inc, Shenzhen, China. Correspondence to: Wenrui Dai <daiwenrui@sjtu.edu.cn>, Xiaopeng Zhang <zxphistory@gmail.com>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

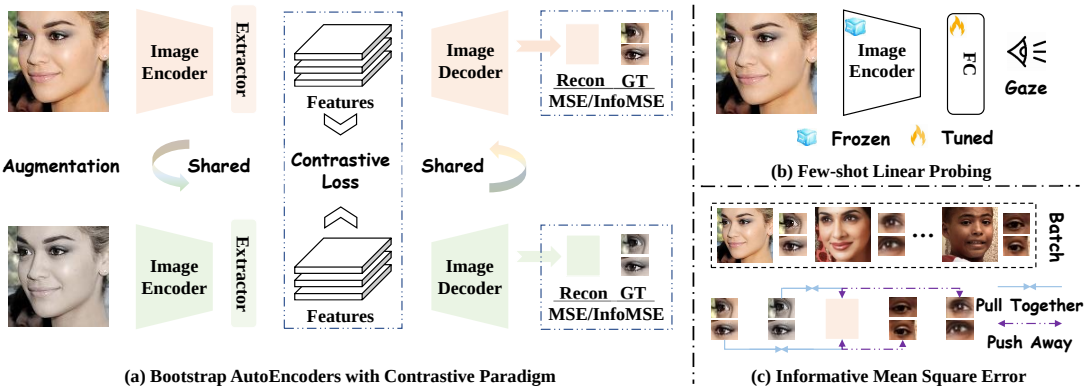


Figure 2: The proposed framework of (a) BeCa, (b) few-shot linear probing using the learned feature, and (c) the proposed InfoMSE. For BeCa, two different augmented images are fed into the Encoder and the output feature is used in three-fold. First, we utilize contrastive loss to regularize the feature. Then, we use a decoder to decode the eye regions. InfoMSE or MSE is used to calculate the loss between the predictions and the ground truth. Finally, we perform linear probing using this feature to get the final gaze direction.

Contrastive (Chen et al., 2020a; Grill et al., 2020; Caron et al., 2021; Zbontar et al., 2021) and generative (He et al., 2022) paradigms are the two mainstream self-supervised learning methods that have achieved impressive performance on various visual tasks, including classification, segmentation, and detection. However, these approaches have encountered challenges when applied to the gaze estimation task, particularly in the context of eye image input. Specifically, contrastive paradigms have been found to be ineffective in self-supervised gaze estimation using eye image input (Sun et al., 2021) (GazeCLR (Jindal & Manduchi, 2023) relies on multi-view video-based gaze images (Park et al., 2020) and the corresponding data augmentation, which are not available in most gaze datasets and are more difficult for practice applications). On the other hand, existing generative self-supervised gaze estimation approaches (Yu & Odobez, 2020; Sun et al., 2021) are limited to eye image input and neglect the valuable gaze information contained in the facial regions. Furthermore, directly expanding these methods to full-face input causes degraded gaze estimation performance, since full-face images contain rich information such as facial details that could overwhelm the model with trivial solutions and produce indistinguishable features for estimating gaze direction.

In this paper, we investigate self-supervised learning for full-face gaze estimation. We first reveal that the vanilla autoencoder only considers reconstruction quality but ignores semantic information for regularizing the learned representation. To address this issue, we propose a novel framework named **B**ootstrap auto-**e**ncoders with **C**ontrastive p**a**radigm (BeCa) to achieve self-supervised learning for full-face gaze estimation by ameliorating the merits of generative and contrastive paradigms. Specifically, we decompose the likelihood function of AutoEncoders and exert an additional contrastive regularization term to facilitate learning distinctive information for identifying gaze direction.

Furthermore, we argue that representation quality and reconstruction quality are not fully correlated, especially when reconstruction quality is good enough. In this way, accurate reconstruction of pixels supervised by the origin image is hardly helpful and could mislead the model to focus on irrelevant details in the image that do not correspond to the gaze direction. As a result, the ground-truth label (image pixel) distribution is not always the optimal label distribution for high-level self-supervised learning. To address this issue, we make a reasonable assumption on the existence of an unknown underlying label distribution, and propose the Informative Mean Square Error (InfoMSE) and develop BeCa-InfoMSE to estimate the mutual information between the input and the suitable label distribution by simultaneously aggregating the predicted label close to its corresponding ground truth and separating the irrelevant pairs.

Experimental results demonstrate that both BeCa and BeCa-InfoMSE achieve superior performance over existing self-supervised approaches in within-dataset and cross-dataset tests. BeCa consistently outperforms existing self-supervised baselines on commonly used gaze datasets, including Columbia, MPII, ETH-Xgaze, and even wild scene *i.e.*, Gaze360. Remarkably, compared with the state-of-the-art eye image self-supervised approach CE (Sun et al., 2021), BeCa achieves 20% lower gaze error on MPII. Furthermore, BeCa-InfoMSE enhances the performance of BeCa with a clear margin of nearly 5% on all the gaze datasets. It is worth mentioning that both BeCa and BeCa-InfoMSE are general to various architectures (*i.e.*, ResNet-18, ResNet-50, and ViT-tiny) and achieve state-of-the-art performance in terms of various kinds of evaluation metrics.

To our best knowledge, we are the first to achieve general self-supervised full-face gaze estimation. The contributions

of this paper are summarized as below.

- We revisit existing generative approaches using the vanilla MSE loss and highlight their limitation in relying solely on reconstruction quality induced by the vanilla MSE loss.
- We propose BeCa that combines the benefits of both MSE loss and the contrastive paradigm for self-supervised full-face gaze estimation. BeCa introduces explicit regularization of semantic information to avoid trivial solutions and facilitate the generative paradigm for full-face images.
- We design a novel InfoMSE loss function to address the issue that unsupervised learning ability and reconstruction quality are not fully correlated, and consequently, develop BeCa-InfoMSE to enhance the quality of learned representation by incorporating information theory into the generative paradigm.

2. Related Work

Appearance-based Gaze Estimation. Deep learning models have facilitated the progress of gaze estimation. Early works (Zhang et al., 2015; 2017a; Cheng et al., 2018) leverage deep neural networks to regress gaze from eye patches since eyes contain rich gaze information. Recently, full-face input is found beneficial to gaze estimation (Zhang et al., 2017b; 2020; Kellnhofer et al., 2019), especially for gaze datasets with large-variance of head pose and gaze, like ETH-XGaze (Zhang et al., 2020) and Gaze360 (Kellnhofer et al., 2019). For instance, (Zhang et al., 2020) uses ResNet-50 to regress gaze from full-face input and obtains a competitive result, especially on gaze generalization benchmark.

Self-supervised Gaze Estimation. Previous self-supervised gaze estimation methods mainly apply to eye images. The main idea is to learn gaze representation by reconstructing the eye with encoder-decoder architecture. (Yu & Odobez, 2020) takes gaze redirection as pretext work for unsupervised gaze learning. Subsequent work CE (Sun et al., 2021) applies the latent-code-swapping mechanism on image pairs, which are the same eye or eye with the same gaze in the same image, and the strategy disentangles the gaze-relevant feature and the appearance-relevant feature. Although GazeCLR (Jindal & Manduchi, 2023) performs full-face self-supervised full-face gaze estimation using multiple encoders and heads on large video-based gaze dataset EVE, these multi-view images are not available in most gaze datasets and are difficult to implement in practice.

Self-supervised Pre-training Contrastive learning methods (Wu et al., 2018; Misra & van der Maaten, 2020; Chen et al., 2020a; He et al., 2020; Caron et al., 2020; 2021; Chen et al., 2022b) have greatly facilitated self-supervised learning in the past few years. Benefiting from the large-scale dataset, the pre-trained model exhibits comparable performance to the supervised counterpart under the linear probing (Xu et al., 2023a) and even better generalization ability in downstream tasks including object detection, instance segmentation, and fine-grained classification (Chen et al., 2020b; Grill et al., 2020). Contrastive learning usually regards the augmented views from the same image as the positive pair and the remaining as negative pairs such that self-supervised learning is achieved by discriminating the positive pair from negative pairs. It is also feasible to simply learn an invariant representation for different augmentations at the risk of collapse (Grill et al., 2020; Caron et al., 2021; Chen & He, 2021). However, contrastive learning cannot be well employed in gaze estimation due to two problems. As mentioned in (Wang et al., 2022), existing contrastive learning methods are based on classification-oriented datasets like ImageNet (Deng et al., 2009) and cannot handle gaze data that is regression-oriented. Furthermore, there is a lack of suitable data augmentation for gaze datasets, and no pseudo-label used in (Wang et al., 2022) is available.

3. Methodology

3.1. Preliminaries

AutoEncoders. Given an image x, an autoencoder first extracts its latent variable z using the encoder f_θ with learnable parameters θ and then recover x from z using an extra decoder g_ϕ with learnable parameters ϕ. We represent the extraction process with the probability distribution $p_\theta(z|x)$ parameterized by θ and the recovery process $p_\phi(y|z)$ by ϕ.

The vanilla auto-encoder deems the same encoder and decoder and maximizes the log-likelihood as:

$$\max \mathbb{E}_{p(x)} \log p_{\phi,\theta}(y|x). \tag{1}$$

Without loss of generality, we assume $p(y|x)$ obeys the Gaussian distribution and rewrite the objective as

$$\max \mathbb{E}_{p(x)} \left[-\|y - \tilde{y}\|_2^2 / \sigma^2 \right], \tag{2}$$

where σ is the variance and $\tilde{y}$ is the ground truth. Note that the objective of the auto-encoder becomes the MSE loss given a constant variance.

Contrastive Learning. Vanilla contrastive learning methods discriminate the positive sample from negative samples through the infoNCE loss (van den Oord et al., 2018). For example, different views of an input image x_i are generated with random data augmentation and encoded into latent variables z_i^k, where i is the index of sample and k is the index of view. Contrastive learning methods maximize the mutual information $\mathbb{E}_{x,z} \log(p(x|z)/p(x))$ between the latent variable z and input image x. Since the mutual information is usually intractable for high-dimensional continuous variables, an approximate InfoNCE loss is alternatively adopted

$$-\log \frac{\exp(\sigma(z_i^k, z_i^q)/\tau)}{\exp(\sigma(z_i^k, z_i^q)/\tau) + \sum_q \sum_j \mathbb{1}_{i \neq j} \exp(\sigma(z_i^k, z_j^q)/\tau)}, \tag{3}$$

where τ is the temperature factor that controls the strength of punishment and $\sigma(\cdot,\cdot)$ is the similarity measurement.

3.2. Bootstrap Auto-Encoders

Consider the dataset $\mathcal{D}$ of N samples $\{x_i, y_i\}_{i=1}^N$, where x_i is an input image and y_i is the label of reconstruction area. For simplicity, we omit the subscript in the rest of this paper. The likelihood $p_\theta(y|x)$ represents the learning procedure that predicts y from x using learnable parameters θ.

Revisiting the Auto-Encoder. As discussed in Section 3.1, vanilla auto-encoders maximize the log-likelihood function $\mathbb{E}_{p(x)} \log p_{\phi,\theta}(y|x)$ but ignore the latent variable z. Here, we deem z as the semantic information containing gaze direction and introduce it into the maximization of $\mathbb{E}_{p(x)} \log p_{\phi,\theta}(y|x)$. Assuming the Markov Chain $y \leftrightarrow z \leftrightarrow x$ for x, y and z, we decompose $p_{\phi,\theta}(y|x)$ as

$$p_{\phi,\theta}(y|x) = \int p_\phi(y|z) p_\theta(z|x) \mathrm{d}z. \tag{4}$$

In Proposition 1, we develop a loss function equivalent to Eq. (4) that relates the auto-encoder to the contrastive loss.

Proposition 1. *Optimizing $\int p_\phi(y|z)p_\theta(z|x)\mathrm{d}z$ is equivalent to optimizing the combination of the contrastive loss between x and z and the MSE loss of y, i.e.,*

$$\mathbb{E}_{x,z} \log \frac{p_\theta(x|z)}{p(x)} + \mathbb{E}_{x,z} \log p_\phi(y|z). \tag{5}$$

Proof. Please refer to the appendix A.1. □

Proposition 1 implies that the latent variable z induces an extra term related to the contrastive loss in the vanilla MSE loss Eq. (2). The first term in Eq. (5) can be viewed as the mutual information between x and z and are approximated using the contrastive (InfoNCE) loss.

Bootstrap auto-encoders with Contrastive paradigm (BeCa). According to Eq. (5), we propose to minimize the combination of contrastive loss and reconstruction loss:

$$-\log \frac{\exp(\mathrm{sim}(z_i, z_k)/\tau)}{\sum_j \exp(\mathrm{sim}(z_i, z_j)/\tau)} + \gamma \cdot \|\tilde{y} - y\|_2^2, \tag{6}$$

where γ is the hyper-parameter and $\mathrm{sim}(\cdot,\cdot)$ is the cosine similarity. Let us define $e_\tau(z_i, z_j) = \exp(\mathrm{sim}(z_i, z_j)/\tau)$ for simplicity. Eq. (6) can be rewritten as

$$-\log \frac{e_\tau(z_i, z_k)}{\sum_j e_\tau(z_i, z_j)} + \gamma \cdot \|\tilde{y} - y\|_2^2. \tag{7}$$

Here, the first term is the contrastive loss that performs as the global regularization to constrain the relationship between the individual samples and the population (*i.e.*, global relationship between the samples at the batch level), while the second term is the MSE loss as the instance-level local objective used in vanilla auto-encoders. In this way, the contrastive paradigm is used to bootstrap the auto-encoders to learn self-supervised gaze representation with explicit global semantic regularization.

3.3. Informative Mean Square Error

BeCa incorporates the contrastive paradigm into AutoEncoders, thereby introducing explicit semantic regularization for pixel reconstruction. While higher reconstruction quality is typically associated with better performance in low-level visual tasks such as super-resolution and denoising, it is not necessarily a reliable indicator of superior performance in high-level vision tasks like self-supervised gaze estimation, which aims to learn semantic information.

Contrary to adopting the ground-truth label distribution $p(y)$, we alternatively assume the underlying label distributions $p_h(y)$ for high-level semantic learning and $p_l(y)$ for low-level visual tasks, respectively. Their likelihoods $p_h(y|x)$ and $p_l(y|x)$ are supposed to share the same conditional distribution $p(x|y)$ and the input distribution $p(x)$.

In this way, directly optimizing the objective in Eq. (1) using the low-level label likelihood $p_l(y)$ is not optimal, while optimizing the objective using the high-level label likelihood $p_h(y)$ is impossible as $p_h(y)$ is unknown. To address this issue, we develop a novel informative mean square error (InfoMSE) to optimize the representation learned from semantic information. Specifically, we focus on the probability ratio $\frac{p(y|x)}{p(y)}$ rather than the likelihood and optimize the mutual information as discussed in Proposition 1. According to the Bayes' Formula, we obtain that

$$\frac{p_h(y|x)}{p_h(y)} = \frac{p(x|y)}{p(x)} = \frac{p_l(y|x)}{p_l(y)}. \tag{8}$$

Then, we develop the InfoMSE as maximizing the mutual information between the input x and the label y as $\mathbb{E}_{x,y} \log \frac{p_h(y|x)}{p_h(y)} = \mathbb{E}_{x,y} \log \frac{p_l(y|x)}{p_l(y)}$. With Eq. (15), we derive the objective of InfoMSE as:

$$\mathbb{E}_{x,y} \log \frac{p_h(y|x)}{p_h(y)} = \mathbb{E}_{x,y} \log \frac{p_l(x|z)}{p_l(x)} + \mathbb{E}_{x,y} \log \frac{p_l(y|\hat{z})}{p_l(y)}. \tag{9}$$

Eq. (9) shows that InfoMSE can be decomposed into two terms, where the first term $\mathbb{E}_{x,y} \log \frac{p_l(x|z)}{p_l(x)}$ is the mutual information between x and z, and the second term $\mathbb{E}_{x,y} \log \frac{p_l(y|\hat{z})}{p_l(y)}$ is the mutual information between label y and gaze feature z. To calculate $\mathbb{E}_{x,y} \log(p_l(y|\hat{z})/p_l(y))$, we first estimate $p_l(y)$ using the Monte Carlo sampling from one mini-batch of M samples $x^1, \cdots, x^M$.

$$p_l(y) = \frac{1}{M} \sum_{j=1}^{M} p_l(y|z^j) \tag{10}$$

where z^j is the latent variable of x^j. Assume that $p_l(y|z)$ is a Gaussian distribution where the expectation is the GT label $\tilde{y}$ with the standard deviation σ, *i.e.*, $p_l(y|z) = \mathcal{N}(y; \tilde{y}, \sigma)$. We have:

$$\mathbb{E}_{x,y} \log \frac{p_l(y|\hat{z})}{p_l(y)} = \mathbb{E}_{x,y} \log \frac{\mathcal{N}(y; \tilde{y}, \sigma)}{\frac{1}{M}\sum_{j=1}^{M} \mathcal{N}(y; \tilde{y}_j, \sigma)} = \log \frac{\exp(-\|y-\tilde{y}\|_2^2/C)}{\frac{1}{M}\sum_{j=1}^{M} \exp(-\|y-\tilde{y}_j\|_2^2/C)}. \quad (11)$$

In this way, We leverage InfoMSE to simultaneously narrow the gap between the predicted label and the corresponding ground truth and discriminate the prediction and ground truth that are not paired.

BeCa With InfoMSE. Furthermore, we propose BeCa-InfoMSE that leverages the proposed InfoMSE rather than MSE in BeCa. Let $e_C(y_i, \tilde{y}_j) = \exp(-\|y_i - \tilde{y}_j\|_2^2/C)$. The objective function for BeCa-InfoMSE is

$$-\log \frac{e_\tau(z_i, z_k)}{\sum_j e_\tau(z_i, z_j)} - \beta \cdot \log \frac{e_C(y_i, \tilde{y}_i)}{\sum_j e_C(y_i, \tilde{y}_j)}, \quad (12)$$

where β is the hyper-parameter.

3.4. Discussion about Full Face and Eye Image

Full-face images contain a wealth of gaze-relevant information, outperforming eye images in supervised scenarios, as demonstrated in recent studies (Bao et al., 2022; Wang et al., 2022; Zhang et al., 2015). However, existing unsupervised gaze approaches overlook this and operate solely on eye images, resulting in inferior performance and reliance on supplementary head pose information. In contrast, our approach capitalizes on the abundant information in full-face images, yielding superior performance even in the absence of head pose information, as shown in Table 4. The utilization of full-face images enables our approach to embrace a fully self-supervised paradigm, hinting at a promising avenue for unsupervised gaze estimation. By harnessing the additional information in full-face images, our approach can enhance the accuracy and robustness of gaze estimation, accommodating a broader range of scenarios and applications.

4. Experiments

4.1. Experimental Settings

Datasets. We perform our self-supervised experiments on 4 gaze datasets: ColumbiaGaze (Smith et al., 2013), MPIIFaceGaze (Zhang et al., 2017b), Gaze360 (Kellnhofer et al., 2019) and ETH-Xgaze (Zhang et al., 2020). The details of these datasets are elaborated in the appendix A.2. Following (Zhang et al., 2018), face normalization is performed on all images according to detected landmarks.

Experimental Details. We employ three different architectures, *i.e.*, ResNet-18, ResNet-50, and ViT-tiny, as the backbones. All experiments are realized using PyTorch(Imambi et al., 2021). Please refer to the appendix A.3 for details.

Linear Probing. We freeze the weights of pre-trained models and train an additional linear regressor using the selected labeled training images following (Sun et al., 2021).

Data Augmentation. Color jitter and grayscale are adopted considering that gaze is sensitive to spatial transformation.

4.2. 100-shot Evaluations

Following the convention in unsupervised learning and previous work (Yu & Odobez, 2020; Sun et al., 2021), we randomly select 100 labeled samples from the training set for either linear probing or few-shot fine-tuning. Then we validate the pre-trained model under the within-dataset and cross-dataset settings. Specifically, We perform 100-shot linear probing and 100-shot fine-tuning under the within-dataset setting, whereas we perform 100-shot linear probing for the cross-dataset setting.

Within-dataset 100-shot linear probing. We assess various existing self-supervised methods for full-face gaze estimation on Columbia, MPII, and Gaze360. The evaluation protocol is linear probing as explained in Section 4.1. Note that BeCa differs from eye-image based methods as head pose is not adopted for end-to-end and fully self-supervised learning. For fair cross-person evaluations, all samples are randomly selected from the training set. Since 100-shot linear probing has not been considered for full-face gaze estimation, we select representative approaches that have achieved notable success in general visual pre-training for comparison, as summarized below.

- **ImageNet-Pretrained.** ImageNet-Pretrained refers to the model pre-trained on ImageNet using supervised learning.
- **AutoEncoder (FRC) and AutoEncoder (ERC).** Facial-patch reconstruction (FRC) reconstructs the whole full-face images using a vanilla auto-encoder architecture, whereas Eye-patch reconstruction (ERC) reconstructs the eye regions using a vanilla auto-encoder architecture.
- **SimCLR (Chen et al., 2020a) and Barlow Twins (Zbontar et al., 2021).** SimCLR and Barlow Twins are two typical contrastive approaches. SimCLR is motivated by the variation between images and performs contrastive learning by enclosing the representations of two augmented views while pulling away other views. In contrast, Barlow Twins focus on the redundancy in feature channels, and learns representations by minimizing their correlations.
- **Masked AutoEncoder (MAE) (He et al., 2022).** MAEs reconstruct masked patches from visible patches using ViTs and achieve impressive performance in general visual tasks.

Table 1: 100-shot linear probing for different self-supervised approaches. †indicates that we use models self-supervised pretrained on Xgaze as the initialization for AutoEncoder (ERC), BeCa and BeCa-InfoMSE on Columbia.

Methods	Backbones	Columbia†	MPII	Gaze360
ImageNet-Pretrained	ResNet-18	13.01±0.2	10.85±0.2	36.70±0.5
AutoEncoder (FRC)	ResNet-18	12.05±0.2	9.21±0.2	30.23±0.4
AutoEncoder (ERC)†	ResNet-18	7.1±0.1	6.60±0.1	26.87±0.4
SimCLR (Chen et al., 2020a)	ResNet-18	13.65±0.2	10.79±0.2	37.29±0.7
Barlow Twins (Zbontar et al., 2021)	ResNet-18	11.77±0.2	9.55±0.1	30.54±0.5
BeCa†	ResNet-18	6.44±0.1	5.76±0.1	22.54±0.3
BeCa-InfoMSE†	ResNet-18	**6.12±0.1**	**5.44±0.1**	**21.75±0.3**
ImageNet-Pretrained	ResNet-50	13.81±0.2	11.34±0.2	38.94±0.6
AutoEncoder (FRC)	ResNet-50	12.49±0.2	9.88±0.2	29.83±0.5
AutoEncoder (ERC)†	ResNet-50	7.46±0.1	7.01±0.1	28.97±0.4
SimCLR (Chen et al., 2020a)	ResNet-50	14.21±0.2	11.28±0.2	38.55±0.6
Barlow Twins (Zbontar et al., 2021)	ResNet-50	12.17±0.2	9.89±0.2	33.17±0.5
BeCa†	ResNet-50	6.52±0.1	5.90±0.1	23.45±0.3
BeCa-InfoMSE†	ResNet-50	**6.37±0.1**	**5.69±0.1**	**22.67±0.3**
ImageNet-Pretrained	ViT-tiny	12.08±0.1	10.22±0.2	37.01±0.5
AutoEncoder (ERC)†	ViT-tiny	7.0±0.1	6.77±0.1	24.25±0.4
Masked AutoEncoder (He et al., 2022)	ViT-tiny	12.1±0.2	8.7±0.1	26.0±0.4
BeCa†	ViT-tiny	6.14±0.1	5.36±0.1	16.23±0.3
BeCa-InfoMSE†	ViT-tiny	**6.03±0.1**	**5.17±0.1**	**15.27±0.4**

Table 2: 100-shot Linear Probing on Xgaze.

Methods	w/ ViT-tiny	w/o ViT-tiny
ImageNet-Pretrained	26.94	42.47
AutoEncoder (ERC)	16.80	19.27
Masker AutoEncoder	28.27	31.13
BeCa	<u>12.82</u>	<u>14.73</u>
BeCa-InfoMSE	**12.17**	**13.54**

Table 3: 100-shot Cross-dataset Linear Probing.

Methods	Backbone	Source	Columbia	MPII
CE (Sun et al., 2021)	ResNet-18	Xgaze	7.76	9.04
BeCa	ResNet-18	Xgaze	7.35	6.51
BeCa-InfoMSE	ResNet-18	Xgaze	**7.08**	**6.26**
BeCa	ViT-tiny	Xgaze	7.23	6.40
BeCa-InfoMSE	ViT-tiny	Xgaze	**7.02**	**6.15**

We employ MAEs to reconstruct the patches randomly masked in full-face images.

As shown in Table 1, BeCa and BeCa-InfoMSE outperform all existing pre-training approaches across different datasets and backbones. Specifically, BeCa achieves 6.44 gaze error, 6.52 gaze error, and 6.14 gaze error on Columbia datasets using ResNet-18, ResNet-50 and ViT-tiny as the backbone respectively. These results achieves average 12% performance improvement compared with AutoEncoder (ERC). For wild gaze dataset, i.e., gaze360, BeCa exhibits over 30% performance improvement when using ViT-tiny as the backbone. Furthermore, InfoMSE consistently improves BeCa across all settings. We also report in Table 2 the results for

Table 4: Comparison with the state-of-the-art self-supervised approaches. We do not use head pose labels.

Methods	Head Pose	Columbia	MPII
(Yu & Odobez, 2020)	-	8.95	-
(Sun et al., 2021)	✓	<u>6.4</u>	7.2
(Sun et al., 2021)	✗	7.1	7.2
(Jindal & Manduchi, 2023)	✗	6.6	6.5
BeCa	✗	<u>6.44</u>	<u>5.76</u>
BeCa-InfoMSE	✗	**6.12**	**5.44**

Table 5: 100-shot fine-tune experiments.

Methods	Backbone	Gaze360
100-shot linear-probing BeCa	ResNet-18	22.54
100-shot Fine-tune BeCa	ResNet-18	**21.81**
100-shot linear-probing BeCa	ResNet-50	23.45
100-shot Fine-tune BeCa	ResNet-50	**22.16**

100-shot linear probing on Xgaze using ViT-tiny as the backbone. Compared with AutoEncoder (ERC), BeCa-InfoMSE reduces the gaze error by 5.73° and achieves a 29.7% gain without the head pose. The gaze error is further reduced to 12.17° by adopting the head pose in BeCa-InfoMSE (*i.e.*, a 27.6% gain over AutoEncoder (ERC)). This result proves that the head pose label can also benefit the proposed approach, despite that we do not use head pose labels in our evaluations in conformity with the self-supervised policy.

Cross-dataset 100-shot linear probing. To evaluate the domain adaptation ability of our approaches, we conduct 100-

Table 6: Linear Probing using the Whole Dataset.

Methods	Backbone	MPII	Gaze360
AutoEncoder (ERC)	ResNet-18	6.22	23.51
AutoEncoder (ERC)	ResNet-50	6.65	27.93
BeCa	ResNet-18	5.33	20.35
BeCa	ResNet-50	5.56	21.82
BeCa-InfoMSE	ResNet-18	**5.19**	**19.95**
BeCa-InfoMSE	ResNet-50	5.37	20.27

Table 7: Whole dataset fine-tuning on MPII

Method	MPII
Baseline (Zhang et al., 2017b)	4.8
RT-Gene (Fischer et al., 2018)	4.8
FAR-Net (Cheng et al., 2020)	4.3
AGE-Net (Biswas et al., 2021)	4.09
L2CS-Net (Abdelrahman et al., 2022)	3.92
BeCa	**3.87**
BeCa-InfoMSE	**3.83**

Table 8: Ablation study on the number of samples used in few-shot linear probing. The backbone is ResNet-18.

Methods	Number	Columbia	MPII
CE (Sun et al., 2021)	50	7.0	8.5
	100	6.4	7.2
	200	6.2	7.3
BeCa-InfoMSE	50	6.75	6.15
	100	6.12	5.44
	200	**6.05**	**5.23**

shot linear probing in a cross-dataset manner. We choose ETH-Xgaze as the source domain dataset, while MPII and Columbia as the target domain dataset. Table 3 shows that BeCa evidently outperforms existing state-of-the-art eye image approaches *i.e.*, CE (Sun et al., 2021) using ResNet-18 as the backbone on both Columbia and MPII datasets. BeCa-InfoMSE further achieves state-of-the-art performance by introducing InfoMSE. Besides, our approach also exhibit impressive using ViT-tiny as the backbone.

Comparison with the state-of-art methods. In Table 4, we compare BeCa and BeCa-InfoMSE with state-of-the-art self-supervised approaches based on eye patches using ResNet-18 as the backbone under the 100-shot linear probing. We do not adopt the head pose, which is used for additional information in (Sun et al., 2021). BeCa and BeCa-InfoMSE are shown to outperform existing approaches on both Columbia and MPII. Besides, BeCa and BeCa-InfoMSE outperform the full-face approach GazeCLR (Jindal & Manduchi, 2023), which is pre-trained on multi-view gaze datasets EVE using multiple encoders and heads.

100-shot fine-tuning. We perform 100-shot fine-tuning on Gaze360 using ResNet-18 and ResNet-50 backbones,

Table 9: Ablation study on the hyper-parameter γ and β

BeCa	Xgaze	BeCa-InfoMSE	Gaze360
$\gamma = 5$	14.73	$\beta = 0.01$	**20.27**
$\gamma = 3$	**14.13**	$\beta = 0.05$	20.44
$\gamma = 1$	14.59	$\beta = 0.1$	21.27
$\gamma = 0.2$	14.63	$\beta = 1$	21.77

Table 10: Ablation on full-face or eye region reconstruction.

Methods	Backbone	ERC	FRC
AutoEncoder	ResNet-18	6.60	9.21
BeCa	ResNet-18	5.76	8.62
BeCa-InfoMSE	ResNet-18	**5.44**	**8.51**
AutoEncoder	ResNet-50	7.01	9.88
BeCa	ResNet-50	5.90	9.31
BeCa-InfoMSE	ResNet-50	**5.69**	**9.15**

Table 11: Experiments using eye images as the input.

Methods	Backbone	Columbia	MPII	Gaze360
Auto-Encoder	ResNet-18	10.6	9.5	-
SimCLR	ResNet-18	10.0	9.8	-
BeCa	ResNet-18	**6.50**	9.38	**36.27**
BeCa-InfoMSE	ResNet-18	6.61	**9.20**	37.27
BeCa	ResNet-50	7.58	9.54	**35.24**
BeCa-InfoMSE	ResNet-50	**7.37**	**9.30**	35.83

as shown in Table 5. Our results demonstrate that few-shot fine-tuning can enhance gaze estimation performance, although the improvement is limited due to the dominant effect of the core self-supervised learning. Nevertheless, these findings suggest that our proposed BeCa model learns effective self-supervised representations.

4.3. Whole-dataset Evaluations

Within-dataset whole-dataset linear probing. We further perform linear probing on the whole dataset. Table 6 shows that, compared with AutoEncoder (ERC), BeCa and BeCa-InfoMSE achieve better performance for both ResNet-18 and ResNet-50 backbones on MPII and Gaze360.

Whole dataset fine-tune experiments. We perform whole-dataset fine-tuning on MPII using BeCa pre-trained on ETH-Xgaze when adopting ResNet-18 as the backbone. Table 7 shows that BeCa and BeCa-InfoMSE outperform the state-of-the-arts for supervised gaze estimation.

4.4. Ablation Studies

Ablation study on InfoMSE. To further investigate the impact of our proposed InfoMSE, we conducted an additional experiment that solely utilizes InfoMSE for self-supervised representation learning. As illustrated in Table 12, we record the 100-shot linear probing performance on MPII

Table 12: Ablation study on the proposed InfoMSE

Method	Generative	Contrastive	InfoMSE	Backbone	MPII	Gaze360	Backbone	MPII	Gaze360
BeCa	✓	✓	×	ResNet18	5.76	22.54	ResNet50	5.90	23.45
BeCa-InfoMSE	✓	✓	✓	ResNet18	5.44	21.75	ResNet50	5.69	22.67
SimCLR	×	✓	×	ResNet18	10.79	37.29	ResNet50	11.28	38.55
AutoEncoder (ERC)	✓	×	×	ResNet18	6.60	26.87	ResNet50	7.01	28.97
Only-InfoMSE	✓	×	✓	ResNet18	5.95	24.93	ResNet50	6.16	24.88

and Gaze360 using ResNet-18 and ResNet-50, respectively. As shown in the table, Only-InfoMSE exhibits a significant performance improvement over AutoEncoder (ERC), which employs the vanilla MSE loss, whereas Only-InfoMSE uses our proposed InfoMSE as the objective function. This suggests that InfoMSE effectively finds a more suitable label distribution and learns a better self-supervised gaze representation. Moreover, BeCa-InfoMSE consistently outperforms BeCa and Only-InfoMSE under all experimental settings. This further supports the notion that InfoMSE enhances self-supervised representation learning by reducing the pursuit of reconstruction quality, and is orthogonal to BeCa.

Number of samples in few-shot linear probing. We evaluate the influence of number of sample in few-shot linear probing in Table 8 and find that the gaze error decreases for BeCa-InfoMSE with the growth of number of samples for few-shot linear probing. BeCa-InfoMSE consistently outperforms CE (Sun et al., 2021) even without using the head pose. This result suggests that the proposed BeCa-InfoMSE learns a better self-supervised gaze representation.

Hyper-parameter γ and β. We explore the influence of hyper-parameter in the loss function. We perform 100-shot linear probing on Xgaze using ViT-tiny as the backbone for BeCa and whole-dataset linear probing on Gaze360 using ResNet-50 as the backbone for BeCa-InfoMSE. Note that BeCa degenerates into SimCLR when $\gamma = 0$. The results are shown in Table 9. We find that the results are relative robust for different hyper-parameter values in BeCa and BeCa-InfoMSE. We set $\gamma = 1$ for BeCa and $\beta = 0.01$ for BeCa-InfoMSE as the default setting.

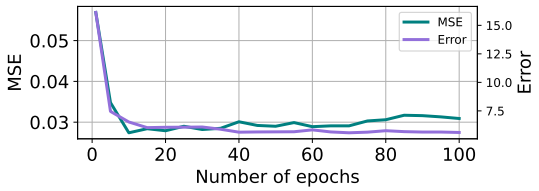
Full-face reconstruction. We have shown that AutoEncoder (FRC) with full-face reconstruction (no contrastive paradigm) is inferior to AutoEncoder (ERC) with eye-region reconstruction in Section 4.2 and Table 1, since full-face reconstruction introduces information unrelated to gaze. Here, we evaluate full-face reconstruction for BeCa and BeCa-InfoMSE on MPII using ResNet-18 and ResNet-50 as backbones. Table 10 shows that full-face reconstruction performs worse than eye-region reconstruction in 100-shot linear probing due to the suppression of useful gaze information by facial information. Furthermore, BeCa and BeCa-InfoMSE outperform AutoEncoder (FRC) as contrastive learning can reduce gaze-unrelated information.

Full face vs. Eye image. We further verify that self-supervised gaze estimation using eye patches is inferior. We evaluate BeCa and BeCa-InfoMSE for eye-patch input on Columbia, MPII, and Gaze360 using ResNet-18 and ResNet50 as the backbones. Table 11 shows that BeCa and BeCa-InfoMSE are superior to vanilla Auto-Encoder and SimCLR (reported by (Sun et al., 2021)). Note that BeCa is designed for full-face gaze estimation and we have addressed why introducing full-face in Section 3.4.

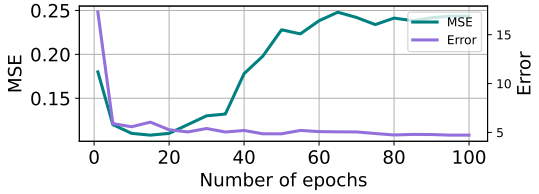
Correlation between reconstruction and representation quality. We visualize the correlation between reconstruction and representation quality for BeCa and BeCa-InfoMSE in Figure 3 to demonstrate that the reconstruction quality is not strongly related to representation quality. We evaluate person 4 in MPII using ResNet-18 and adopt MSE on test data to measure reconstruction quality and gaze prediction error on test data for representation quality. After 100 epochs of training, BeCa achieves a 5.58 gaze error with 0.029 MSE loss, while BeCa-InfoMSE achieves a 4.72 gaze error with 0.24 MSE loss. These results suggest that detailed reconstruction does not contribute to self-supervised gaze estimation and may even impair the extraction of gaze information, as we discussed in Sec. 3.3. Moreover, we observe from Figure 3b that when using BeCa-InfoMSE, the MSE loss, which represents the reconstruction quality, decreases in the early training period but eventually rises as the MSE loss decreases to a certain extent. However, the corresponding gaze prediction error continues to decrease. This result further indicates that reconstruction quality is not positively correlated with semantic representation, and our BeCa-InfoMSE finds a more suitable label distribution.

Extension Experiments on Feature Visualization. We visualize the distribution of features learned on MPII with t-SNE (van der Maaten & Hinton, 2008) to validate the effectiveness of learning a good representation. We use ResNet-18 as the backbone. Figure 4 shows that the ImageNet-pretrained model learns chaotic and disorder features, while BeCa-InfoMSE exhibits a good feature arrangement.

Extensive Experiment on Gaze360. Table 13 illustrates that our approach considers a broader range of usage scenarios and is effective in these complex scenarios, we test the current SOTA approach i.e., GAZECLR on Gaze360. To ensure a fair comparison, we first used the official checkpoint and tested MPII performance (achieving 6.41 in our

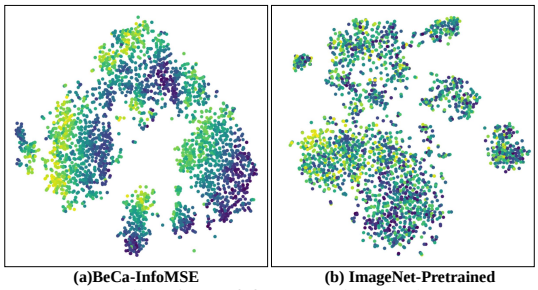


(a) BeCa



(b) BeCa-InfoMSE

Figure 3: Visualization of the correlations between reconstruction and representation quality. Here the MSE loss represents the reconstruction (lower means better reconstruction) while the Error indicates the gaze error (lower means better representation).



(a)BeCa-InfoMSE (b) ImageNet-Pretrained

Figure 4: Visualization of features learned on MPII using (a) BeCa-InfoMSE and (b) ImageNet-pretrained ResNet-18. Different gaze labels are marked by different colors.

Table 13: Experiment on Gaze360.

Method	Columbia	MPII	Gaze360	XGAZE
GAZECLR	6.6	6.5	37.07	29.91
BeCa	6.44	5.76	22.54	19.40
BeCa-InfoMSE	6.12	5.44	21.75	18.97

test versus 6.5 reported in GAZECLR's original article). Then, we evaluated GAZECLR with the official checkpoint on Gaze360 and XGAZE using ResNet-18 as the backbone. We find that GAZECLR achieves very poor performance compared to our BeCa and BeCa-InfoMSE. This indicates that our approach is more practical and better at learning self-supervised gaze representation.

Extensive Experiments for pretraining in Columbia. For the pre-training in Columbia, we perform two experiments to verify the effectiveness of our proposed approaches. Firstly, we perform within-dataset 100-shot linear probing with ResNet-18 as the backbone on Columbia using the model pre-trained on XGAZE with BeCa as the initialization of AutoEncoder (ERC) (the second line in Table 14).

Table 14: Whether load BeCa's pretrained XGAZE.

Method	Pretrained	Load BeCa	Columbia
AutoEncoder (ERC)	23.25	×	7.10
AutoEncoder (ERC)	19.40	✓	6.82
BeCa	19.40	✓	6.44

Table 15: Experiment on whether pretraining on XGAZE.

Method	Backbone	pretrained	$\mathcal{D}_C \rightarrow \mathcal{D}_M$
AutoEncoder (ERC)	ResNet18	×	8.40
BeCa	ResNet18	×	6.15
BeCa-InfoMSE	ResNet18	×	5.06
AutoEncoder (ERC)	ResNet18	✓	5.33
BeCa	ResNet18	✓	4.62
BeCa-InfoMSE	ResNet18	✓	4.47

We compare this with the model pre-trained on XGAZE with AutoEncoder (ERC) as the initialization (the first line). The results show that the second line exhibits better linear probing performance compared to the first line, indicating that BeCa learns better gaze representation and the pre-training can facilitate the downstream feature learning. Furthermore, even using the same pre-training, BeCa outperforms AutoEncoder (ERC) with a clear improvement (6.44 vs. 6.82). This result suggests that BeCa can further help the downstream dataset learn the self-supervised representation.

Furthermore, we conduct cross-dataset 100-shot linear probing experiments on Columbia ($\mathcal{D}_C$) $\rightarrow$ MPII ($\mathcal{D}_M$) using ResNet-18 as the backbone, comparing the performance of our proposed approaches with and without pre-training. The results are presented in Table 15. From the first three lines of the table, it is evident that when no pre-training is used, BeCa and BeCa-InfoMSE significantly outperform the baseline AutoEncoder (ERC) by over 25%. Additionally, BeCa-InfoMSE shows a clear improvement over BeCa. These results suggest that even without pre-training, our method is capable of learning better self-supervised gaze representations and vastly outperforms the baseline.

5. Conclusion and Discussion

In this paper, we revisit the vanilla AutoEncoder and find that the objective function only focuses on reconstruction quality, neglecting semantic regularization. To address this issue, we propose BeCa, which leverages the contrastive paradigm to explicitly regularize the gaze self-supervised representation. Furthermore, we claim that representation learning ability and reconstruction quality are not fully correlated, especially when reconstruction quality is good enough. Thus, we propose the InfoMSE to estimate the mutual information between the semantic information and the reconstruction label y. Experimental results demonstrate the effectiveness of our work under both within-dataset and cross-dataset manner using different kinds of backbones.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under Grant 62125109, Grant 61931023, Grant 61932022, Grant 62371288, Grant 62320106003, Grant 62301299, Grant T2122024, Grant 62120106007, Grant 62250055.

Impact Statement

This paper presents work whose goal is to advance the field of self-supervised gaze estimation. Since faces and privacy-sensitive gazes are involved, care needs to be taken to protect these private data.

References

Abdelrahman, A. A., Hempel, T., Khalifa, A., and Al-Hamadi, A. L2CS-Net: Fine-grained gaze estimation in unconstrained environments. *arXiv preprint arXiv:2203.03339*, 2022.

Ba, J. L., Kiros, J. R., and Hinton, G. E. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

Bao, Y., Liu, Y., Wang, H., and Lu, F. Generalizing gaze estimation with rotation consistency. In *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 4207–4216, 2022.

Biswas, P. et al. Appearance-based gaze estimation using attention and difference mechanism. In *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 3143–3152, 2021.

Burova, A., Mäkelä, J., Hakulinen, J., Keskinen, T., Heinonen, H., Siltanen, S., and Turunen, M. Utilizing vr and gaze tracking to develop ar solutions for industrial maintenance. In *Proc. 2020 CHI Conf. Human Factors Comput. Syst.*, pp. 1–13, 2020.

Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., and Joulin, A. Unsupervised learning of visual features by contrasting cluster assignments. In *Adv. Neural Inf. Process. Syst. 33*, pp. 9912–9924, 2020.

Caron, M. et al. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 9630–9640, 2021.

Chen, S.-Y., Lai, Y.-K., Xia, S., Rosin, P., and Gao, L. 3D face reconstruction and gaze tracking in the HMD for virtual interaction. *IEEE Trans. Multimedia*, 25:3166–3179, 2022a.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In *Proc. 37th Int. Conf. Mach. Learn.*, pp. 1597–1607, 2020a.

Chen, X. and He, K. Exploring simple Siamese representation learning. In *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 15750–15758, 2021.

Chen, X., Fan, H., Girshick, R. B., and He, K. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020b.

Chen, X. et al. Context autoencoder for self-supervised representation learning. *arXiv preprint arXiv:2202.03026*, 2022b.

Cheng, Y. and Lu, F. DVGaze: Dual-view gaze estimation. In *2023 IEEE/CVF Int. Conf. Comput. Vis.*, pp. 20632–20641, 2023.

Cheng, Y., Lu, F., and Zhang, X. Appearance-based gaze estimation via evaluation-guided asymmetric regression. In *15th Eur. Conf. Comput. Vis. (ECCV)*, pp. 100–115, 2018.

Cheng, Y., Zhang, X., Lu, F., and Sato, Y. Gaze estimation by exploring two-eye asymmetry. *IEEE Trans. Image Process.*, 29:5259–5272, 2020.

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

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 248–255, 2009.

Dosovitskiy, A. et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *8th Int. Conf. Learn. Rep.*, 2020.

Fischer, T., Chang, H. J., and Demiris, Y. RT-GENE: Real-time eye gaze estimation in natural environments. In *15th Eur. Conf. Comput. Vis. (ECCV)*, pp. 334–352, 2018.

Fletcher, L. and Zelinsky, A. Driver inattention detection based on eye gaze—road event correlation. *Int. J. Rob. Res.*, 28(6):774–801, 2009.

Gerber, M. A., Schroeter, R., Xiaomeng, L., and Elhenawy, M. Self-interruptions of non-driving related tasks in automated vehicles: Mobile vs head-up display. In *Proc. 2020 CHI Conf. Human Factors Comput. Syst.*, pp. 1–9, 2020.

Grill, J.-B. et al. Bootstrap your own latent: A new approach to self-supervised learning. In *Adv. Neural Inf. Process. Syst. 33*, pp. 21271–21284, 2020.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 9729–9738, 2020.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 16000–16009, 2022.

Hutchinson, T. E., White, K. P., Martin, W. N., Reichert, K. C., and Frey, L. A. Human-computer interaction using eye-gaze input. *IEEE Trans. Syst., Man, Cybern.*, 19(6): 1527–1534, 1989.

Imambi, S., Prakash, K. B., and Kanagachidambaresan, G. Pytorch. *Programming with TensorFlow: Solution for Edge Computing Applications*, pp. 87–104, 2021.

Ioffe, S. and Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. 32nd Int. Conf. Mach. Learn.*, pp. 448–456, 2015.

Jindal, S. and Manduchi, R. Contrastive representation learning for gaze estimation. In *Proc. 1st Gaze Meets ML Workshop*, pp. 37–49, 2023.

Kellnhofer, P., Recasens, A., Stent, S., Matusik, W., and Torralba, A. Gaze360: Physically unconstrained gaze estimation in the wild. In *2019 IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 6912–6921, 2019.

Kim, K.-N. and Ramakrishna, R. Vision-based eye-gaze tracking for human computer interface. In *1999 IEEE Int. Conf. Syst., Man, Cybern.*, pp. 324–329, 1999.

Konrad, R., Angelopoulos, A., and Wetzstein, G. Gaze-contingent ocular parallax rendering for virtual reality. *ACM Trans. Graph. (TOG)*, 39(2):1–12, 2020.

Lei, Y., He, S., Khamis, M., and Ye, J. An end-to-end review of gaze estimation and its interactive applications on handheld mobile devices. *ACM Comput. Surv.*, 56(2): 1–38, 2023.

Liu, G., Yu, Y., Mora, K. A. F., and Odobez, J.-M. A differential approach for gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(3):1092–1099, 2021.

Liu, H. and Qin, H. Perceptual self-position estimation based on gaze tracking in virtual reality. *Virtual Reality*, pp. 1–10, 2022.

Lu, F., Sugano, Y., Okabe, T., and Sato, Y. Adaptive linear regression for appearance-based gaze estimation. *IEEE Tran. Pattern Anal. Mach. Intell.*, 36(10):2033–2046, 2014.

Martin, S., Vora, S., Yuen, K., and Trivedi, M. M. Dynamics of driver's gaze: Explorations in behavior modeling and maneuver prediction. *IEEE Trans. Intell. Veh.*, 3(2):141–150, 2018.

Misra, I. and van der Maaten, L. Self-supervised learning of pretext-invariant representations. In *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 6707–6717, 2020.

Murthy, L. and Biswas, P. Appearance-based gaze estimation using attention and difference mechanism. In *2021 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, pp. 3137–3146, 2021.

Pal, A., Mondal, S., and Christensen, H. I. "looking at the right stuff"-guided semantic-gaze for autonomous driving. In *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 11883–11892, 2020.

Park, S., Aksan, E., Zhang, X., and Hilliges, O. Towards end-to-end video-based eye-tracking. In *16th Eur. Conf. Comput. Vis.*, pp. 747–763, 2020.

Paszke, A. et al. PyTorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inf. Process. Syst. 32*, pp. 8024–8035. 2019.

Patney, A., Salvi, M., Kim, J., Kaplanyan, A., Wyman, C., Benty, N., Luebke, D., and Lefohn, A. Towards foveated rendering for gaze-tracked virtual reality. *ACM Trans. Graph. (TOG)*, 35(6):1–12, 2016.

Shi, P., Billeter, M., and Eisemann, E. Salientgaze: Saliency-based gaze correction in virtual reality. *Computers & Graphics*, 91:83–94, 2020.

Smith, B., Yin, Q., Feiner, S., and Nayar, S. Gaze locking: Passive eye contact detection for human?object interaction. In *ACM Symp. User Interface Software Technol. (UIST)*, pp. 271–280, 2013.

Sun, Y., Zeng, J., Shan, S., and Chen, X. Cross-encoder for unsupervised gaze representation learning. In *2021 IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 3702–3711, 2021.

Surakka, V., Illi, M., and Isokoski, P. Gazing and frowning as a new human–computer interaction technique. *ACM Trans. Appl. Percept. (TAP)*, 1(1):40–56, 2004.

van den Oord, A., Li, Y., and Vinyals, O. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748v2*, 2018.

van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.*, 9(86):2579–2605, 2008.

Vaswani, A. et al. Attention is all you need. In *Adv. Neural Inf. Process. Syst. 30*, pp. 5998–6008, 2017.

Wang, Y. et al. Contrastive regression for domain adaptation on gaze estimation. In *2022 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 19376–19385, 2022.

Wu, Z., Xiong, Y., Yu, S. X., and Lin, D. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, pp. 3733–3742, 2018.

Xu, H. et al. Seed the views: Hierarchical semantic alignment for contrastive representation learning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(3):3753–3767, 2023a.

Xu, M., Wang, H., and Lu, F. Learning a generalized gaze estimator from gaze-consistent feature. In *Proc. 37th AAAI Conf. Artif. Intell.*, pp. 3027–3035, 2023b.

Yu, Y. and Odobez, J.-M. Unsupervised representation learning for gaze estimation. In *2020 IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pp. 7314–7324, 2020.

Zbontar, J., Jing, L., Misra, I., LeCun, Y., and Deny, S. Barlow twins: Self-supervised learning via redundancy reduction. In *Proc. 38th Int. Conf. Mach. Learn.*, pp. 12310–12320, 2021.

Zhang, X., Sugano, Y., Fritz, M., and Bulling, A. Appearance-based gaze estimation in the wild. In *2015 IEEE Conf. Comput. Vis. Pattern Recognit.*, pp. 4511–4520, 2015.

Zhang, X., Sugano, Y., Fritz, M., and Bulling, A. MPIIGaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(1): 162–175, 2017a.

Zhang, X., Sugano, Y., Fritz, M., and Bulling, A. It's written all over your face: Full-face appearance-based gaze estimation. In *2017 IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, pp. 51–60, 2017b.

Zhang, X., Sugano, Y., and Bulling, A. Revisiting data normalization for appearance-based gaze estimation. In *Proc. 2018 ACM Symp. Eye Tracking Res. Appl.*, 2018.

Zhang, X., Park, S., Beeler, T., Bradley, D., Tang, S., and Hilliges, O. ETH-XGaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *16th Eur. Conf. Comput. Vis. (ECCV)*, pp. 365–381, 2020.

A. Appendix.

A.1. Proof of the Proposition

Proposition 1. *Optimizing* $\int p_\phi(y|z)p_\theta(z|x)\mathrm{d}z$ *is equivalent to optimizing the combination of the contrastive loss between* x *and* z *and the MSE loss of* y*, i.e.,*

$$\mathbb{E}_{x,z}\log\frac{p_\theta(x|z)}{p(x)}+\mathbb{E}_{x,z}\log p_\phi(y|z). \tag{13}$$

Proof. According to Bayes' Theorem, $p_\theta(z|x) = p_\theta(x|z)p(z)/p(x)$. Thus, the integral $\int p_\phi(y|z)p_\theta(z|x)\mathrm{d}z$ can be rewritten as:

$$\int p_\phi(y|z)p_\theta(z|x)\mathrm{d}z=\int p_\phi(y|z)\frac{p_\theta(x|z)p(z)}{p(x)}\mathrm{d}z. \tag{14}$$

Eq. (14) can be estimated with Monte Carlo sampling as:

$$\int p_\phi(y|z)\frac{p_\theta(x|z)p(z)}{p(x)}\mathrm{d}z=p_\phi(y|\hat{z})\frac{p_\theta(x|\hat{z})}{p(x)}, \tag{15}$$

where $\hat{z}$ is the feature sampled from $p(z|x)$. Then the log-likelihood function is

$$\mathbb{E}_x\log\left[p_\phi(y|\hat{z})\frac{p_\theta(x|\hat{z})}{p(x)}\right]=\mathbb{E}_x\log\frac{p_\theta(x|\hat{z})}{p(x)}+\mathbb{E}_x\log p_\phi(y|\hat{z}). \tag{16}$$

Therefore, we draw Proposition 1. □

A.2. Experimental Settings

Datasets. We perform our self-supervised experiments on 4 gaze datasets: ColumbiaGaze (Smith et al., 2013), MPIIFaceGaze (Zhang et al., 2017b), Gaze360 (Kellnhofer et al., 2019) and ETH-Xgaze (Zhang et al., 2020). The details of these datasets are elaborated below.

- **ColumbiaGaze** (Columbia) consists of 5.8K images from 56 subjects. Following the convention, a 5-fold evaluation protocol is adopted for Columbia.
- **MPIIFaceGaze** (MPII) is collected from 15 subjects in front of laptops. 3,000 images per subject are evaluated with a leave-one-out evaluation protocol.
- **Gaze360** contains labeled full-face images with a wide-range head pose. Following (Cheng et al., 2021), we remove images without faces and use the left 84,902 images as the training set. We follow the official train-test splitting and use 16,031 test images for evaluation.
- **ETH-Xgaze** (Xgaze) is collected with 18 digital SLR cameras from 110 participants in laboratory environments. We use 756,540 images from 80 subjects as the training set and an official test set consisting of 150k images for evaluating self-supervised learning performance.

Following (Zhang et al., 2018), face normalization is performed on all images according to detected landmarks.

Table 16: ResNet architecture

Config	Value
Details for ResNet-18	
image size	224
output-dim	512
Params	11.4 M
Details for ResNet-50	
image size	224
output-dim	2048
Params	23.7 M

Table 17: ViT-tiny architecture

Config	Value
image size	224
patch size	16
encoder embedding dim	192
encoder depth	12
encoder num heads	3
decoder embedding dim	192
decoder depth	4
decoder num heads	3
Params	5.49M

A.3. Implement Details

ResNet architecture. For ResNet-18 and ResNet-50 backbone, we utilize the official model and the details are illustrated in Table 16. After the backbone, we employ a linear layer to output 16-dim feature as the feature used for linear probing. Then an additional linear layer is employed with $16 \rightarrow 128$ dim and output 128-dim feature, which is used for contrastive learning. For the decoder, we first use a pre-linear layer to output a $512 * 1 * 1$ feature and then utilize 7 blocks for up-sampling and each block is formed as ConvTranspose2d-BatchNorm2d-LeakyReLU. The hidden dim for each block is $512 \rightarrow 256, 256 \rightarrow 128, 128 \rightarrow 64, 64 \rightarrow 32, 32 \rightarrow 16, 16 \rightarrow 8, 8 \rightarrow 4$. Then for the output feature with a shape of $[4, 72, 60]$, we then employ a vanilla 2D convolution to reshape this feature with the same shape as the target eye region ([3,72,60]). The params for the decoder is 1.4M for ResNet-18 and 1.7M for ResNet-50.

ViT-tiny architecture. Following MAE (He et al., 2022), We use the standard ViT architecture (Dosovitskiy et al., 2020), which has a stack of Transformer blocks (Vaswani et al., 2017). To keep the same as the MAE, we add the LayerNorm (Ba et al., 2016) to the end of the Encoder. As the decoder is much smaller than the encoder, a linear projection layer is adopted after the encoder to match the dimension of the encoder features and decoder features as well. The details of the ViT-tiny are shown in Table 17.

Self-supervised Pre-training setting. We implement the codes with the Pytorch (Paszke et al., 2019) framework and use 4 Nvidia-V100 GPUs for training. An AdamW optimizer and a cosine decay learning rate schedule are used with the initial learning rate settled as 4×10^{-4} for ResNet and 1.5×10^{-4} for ViT-tiny. A 0.05 weight-decay is also employed and we warm up the training process with 10 epochs and then train the model for 190 epochs (The total epochs are 200). We use the ground-truth eye landmarks to capture eye images following (Cheng et al., 2021). We employ random color jitter and gray transform as data augmentation for contrastive paradigms. All the models not only our approaches are initialized from weights pre-trained on ImageNet. A linear learning-rate scaling rule: $lr = baselr \times batchsize/256$ is used for distributed multi-Gpu training.

Few-shot linear-probing setting. For few-shot linear probing, the model is frozen while only an additional linear regressor is trained. The loss function of linear probing is **L1 Loss** between the predicted gaze angle and target gaze. The batch size is set to 28 and the base learning rate is 4×10^{-4}. Following MAE, an extra Batch Normalization (BN) layer (Ioffe & Szegedy, 2015) without affine transformation is employed before the linear regressor to stable the training.

Few-shot fine-tune setting. The setting of few-shot fine-tuning is similar to few-shot linear-probing. The major difference is that we train the whole model rather than only the linear regressor.

Whole dataset linear-probing setting. The setting of whole-dataset linear-probing is similar to that of few-shot linear-probing. The difference is that we use $batchsize = 128$ and linear-probing for 20 epochs.

Whole dataset Fine-tune setting. For whole dataset fine-tune, we use the Adam optimizer along with a cosine learning rate scheduler. We adopt 20 epochs training, where the first 3 epochs are used for warm-up training. We use the feature generated from the backbone and then we use a one-layer linear classifier to output 2-dim gaze angle prediction. With this setting, we hope to make a fair comparison with the supervised gaze estimation.

A.4. Limitation and Future work

As we discussed in the main text, we need to pre-trained on ETH-Xgaze for a good initialization for Columbia few-shot linear probing. This is because Columbia has too few samples to learn a good representation for full-face gaze estimation. Besides, our approaches still need eye landmark information like eye image approaches. Learning self-supervised gaze representations from full faces without any labels remains challenging in future research.