Contrastive Representation Learning for Gaze Estimation

Swati Jindal
University of California, Santa Cruz
swjindal@ucsc.edu

Roberto Manduchi
University of California, Santa Cruz
manduchi@soe.ucsc.edu

Abstract

Self-supervised learning (SSL) has become prevalent for learning representations in computer vision. Notably, SSL exploits contrastive learning to encourage visual representations to be invariant under various image transformations. The task of gaze estimation, on the other hand, demands not just invariance to various appearances but also equivariance to the geometric transformations. In this work, we propose a simple contrastive representation learning framework for gaze estimation, named *Gaze Contrastive Learning (GazeCLR)*. *GazeCLR* exploits multi-view data to promote equivariance and relies on selected data augmentation techniques that do not alter gaze directions for invariance learning. Our experiments demonstrate the effectiveness of *GazeCLR* for several settings of the gaze estimation task. Particularly, our results show that *GazeCLR* improves the performance of cross-domain gaze estimation and yields as high as 17.2% relative improvement. Moreover, the *GazeCLR* framework is competitive with state-of-the-art representation learning methods for few-shot evaluation. The code and pre-trained models are available at https://github.com/jswati31/gazeclr.

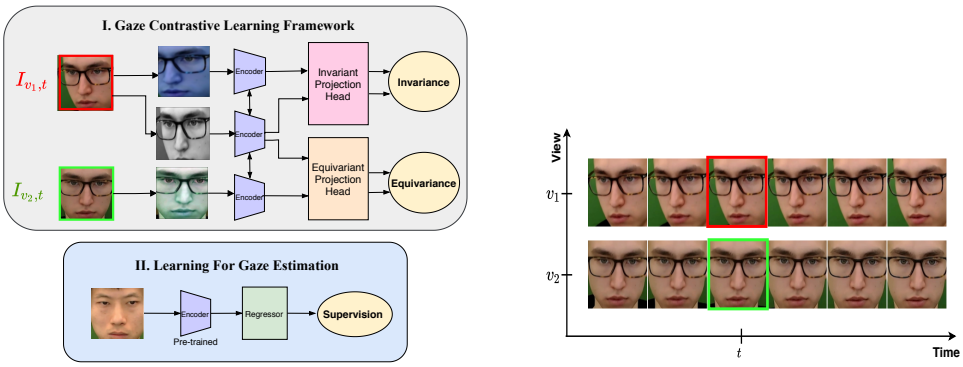
1 Introduction

Gaze represents the focus of human attention and serves as an essential cue for non-verbal communication. While specialized gaze trackers can accurately measure a user's gaze direction, there is substantial interest in gaze estimation using regular cameras. Although, learning gaze estimation models from images is challenging and needs to transcend multiple "nuisance" characteristics such as facial features or head orientation to estimate gaze accurately.

In recent years, deep learning [1, 2, 3] has shown promising results for gaze estimation. In part, this success stems from the availability of large-scale annotated datasets. As a result, valuable datasets must contain a wide range of gaze directions, appearances, and head poses, which is a laborious and time-consuming procedure. Also, gaze annotations are difficult to obtain, which makes the creation of large, representative datasets challenging [4]. Therefore, methods that facilitate training with limited gaze annotations are highly desirable.

Self-supervised learning (SSL) has gained tremendous success over the past few years and emerged as a powerful tool for reducing over-reliance on human annotations [5, 6, 7]. Following a generally accepted paradigm, we consider a pre-training stage that requires no labels, followed by a fine-tuning stage using a relatively small number of labeled samples. SSL is an effective approach for pre-training, where semantically meaningful representations are learned that can be seamlessly adapted during fine-tuning stage [8, 9, 10]. Specifically, a good pre-training would ensure that the embeddings for images associated with the same gaze direction are neighbors in the feature space, regardless of other non-relevant factors such as appearance. Arguably, this could accelerate the job of fine-tuning, possibly reducing the number of required labeled samples.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).



(a) Two-stages of our framework

(b) Formation of positive pairs

Figure 1: **Overall idea.** (a) The proposed two-stage learning framework for gaze estimation. Stage-I shows Gaze Contrastive Learning (*GazeCLR*) framework trained using only unlabeled data and learns both *invariance* and *equivariance* properties. In Stage-II, the pre-trained encoder is employed for gaze estimation task with small labeled data. (b) Two images (shown in **red** and **green**) captured at same time with different camera views are used to create both invariant and equivariant positive pairs.

In this work, for SSL pre-training, we focus on *contrastive representation learning* (CRL), which aims to map "positive" pair samples to embeddings that are close to each other, while mapping "negative" pairs apart from each other [11]. A popular approach is to generate pairs by applying two different transformations (or augmentations) to an input image forming a positive pair, and different images forming negative pairs. This method encourages invariance in representations w.r.t. similar types of transformations, where these transformations are assumed to model "nuisance" effects.

However, obtaining the necessary and sufficient set of "positive" and "negative" pairs remains a non-trivial and unanswered challenge for a given task. This work attempts to answer this question for gaze estimation. Recent CRL-based methods encourage the representations to be invariant to *any* image transformation, many of which are not suitable for gaze estimation. For example, geometry-based image transformations (such as rotation) will change the gaze direction. In contrast, it is beneficial to have invariance to appearance, e.g., a person's identity, background, etc.

In this paper, we propose *Gaze Contrastive Learning* (or *GazeCLR*) framework – a simple CRL-based unsupervised pre-training approach for gaze estimation, i.e., a pre-training method requiring no gaze label data. In detail, our approach relies on *invariance* to image transforms (e.g., color jitter) that do not alter gaze direction, and *equivariance* to camera viewpoint, which requires additional information of multi-view geometry, i.e., images of the same person should be obtained at the same time by two or more cameras from different locations.

For learning *equivariance*, we leverage the fact that in *a common reference system*, two or more synchronous images of the same person from different camera viewpoints are associated with the same gaze direction. The knowledge of the relative pose of each camera to the *common reference system* provides the relation of gaze directions defined in the respective camera space. In other words, gaze direction has an equivariant relationship to camera viewpoints. We claim that the requirement of using multiple cameras may be less onerous than obtaining gaze annotations for each image.

We use an existing multi-view gaze dataset (EVE [12]) which provides video sequences captured from four calibrated and synchronized cameras and contains gaze annotations, which are obtained using a gaze tracking device [13]. We neglect labels during pre-training and use them only for fine-tuning and evaluation. Observe that, the relative camera pose information available with the EVE dataset is used *only* during pre-training stage. Figure 1 presents an overview of the proposed idea.

To evaluate the *GazeCLR*, we perform self-supervised pre-training using the EVE dataset and transfer the learned representations for the gaze estimation task in various evaluation settings. We demonstrate the effectiveness of representations by showing that the proposed method achieves superior performance on both within-dataset and cross-dataset (MPIIGaze [2] and Columbia [14]) evaluations by using only a small number of labeled samples for fine-tuning. Our major contributions are summarized as follows:

1. We propose a simple contrastive learning method for gaze estimation that relies on the *observation* that gaze direction is *invariant* under selected appearance transformations and *equivariant* to any two camera viewpoints.
2. We also argue to learn equivariant representations by taking advantage of the multi-view data that can be seamlessly collected using multiple cameras.
3. Our empirical evaluations show that *GazeCLR* yields improvements for various settings of gaze estimation and is competitive with existing supervised [15] and unsupervised state-of-the-art gaze representation learning methods [16, 17].

2 Proposed Method

2.1 Stage-I: Gaze Contrastive Learning (*GazeCLR*) Framework

GazeCLR is a framework to train an *encoder* that learns embeddings to induce desired set of invariance and equivariance for the gaze estimation task. As stated earlier, the key intuition of *GazeCLR* is that we want to enforce invariance using selected appearance transformations (e.g., color jitter) and equivariance using synchronous images of the same person captured from multiple camera viewpoints. Similar to previous SSL approaches [6, 18], we rely on the normalized temperature-scaled cross-entropy loss (NT-Xent)[6] to encourage invariance or equivariance by maximizing the agreement between positive pairs and disagreement between the negative pairs. In particular, we devise two variants of NT-Xent loss, namely, $\mathcal{L}^I$ for invariance and $\mathcal{L}^E$ for equivariance.

The *GazeCLR* framework has three sub-modules: a CNN-based *encoder* and two *projection heads* based on MLP layers, as illustrated in Figure 2. The output of the encoder branches out into different *projection heads* depending on the type of input positive pair. To abide by the invariance for gaze direction, we consider augmentations based on only *appearance* transformations denoted as $\mathcal{A}$.

Let $\{I_{v_i,t}\}_{i=1}^{|V|}$ be the synchronous frames for timestamp t coming from different camera views (i.e., $\{v_i\}_{i=1}^{|V|}$), then we create following positive pairs:

1. *Single-view positive pairs:* We apply two randomly sampled augmentations from $\mathcal{A}$ to create a single-view positive pair. Specifically, for any image $I_{v_i,t}$, at a given timestamp t and view v_i, we sample two augmentations a and a' from $\mathcal{A}$ and then $(a(I_{v_i,t}), a'(I_{v_i,t}))$ forms a positive pair to learn invariance. The left branch of Figure 2 shows one such positive pair for view v_1.
2. *Multi-view positive pairs:* We consider all unique pairs of camera viewpoints from the same timestamp t and apply random augmentations from $\mathcal{A}$, i.e., $\{(a_i(I_{v_i,t}), a_j(I_{v_j,t})) \mid i, j \in \{1, \ldots, |V|\} \mid i \neq j\}$. The corresponding outputs from the encoder are passed through projection head p_2 and multiplied by an appropriate rotation matrix to learn equivariance.

Next, to construct negative pairs, we do not sample them explicitly but use all other samples in the mini-batch as negative examples, similar to Chen et al. [6]. The exact formulation of both loss functions $\mathcal{L}^I$ and $\mathcal{L}^E$ is described below. For brevity, we omit t from $I_{v_i,t}$ and augmentation a in the following subsections.

Single-View Learning. Recall, the goal of *single-view* learning is to induce invariance amongst representations under various appearance transformations. Let $v_i \in V$ be any view and $b \in [1, \ldots, B]$ be the batch index. Given a batch size of B, we apply two augmentations to each sample in the batch yielding $2B$ augmented images, and for each sample, we have one positive pair and $(2B-1)$ negative pairs stemming from remaining samples in the batch. Our encoder E extracts representations for all $2B$ augmented images, which are further mapped by projection head $p_1(\cdot)$ yielding embeddings $(\{z_{v_i}^b, z_{v_i}'^b\}_{b=1}^B)$. With above notations, for any view v_i, the proposed invariance loss function $\mathcal{L}^I$ associated with a positive pair $(z_{v_i}^b, z_{v_i}'^b)$ can be given as follows:

$$\mathcal{L}^I(z_{v_i}^b, z_{v_i}'^b) = -\,\texttt{log}\,\frac{\texttt{sim}(z_{v_i}^b, z_{v_i}'^b)}{\sum_{l=1}^B 1_{l\neq b}\,\texttt{sim}(z_{v_i}^b, z_{v_i}^l) + \sum_{l=1}^B \texttt{sim}(z_{v_i}^b, z_{v_i}'^l)} \tag{1}$$

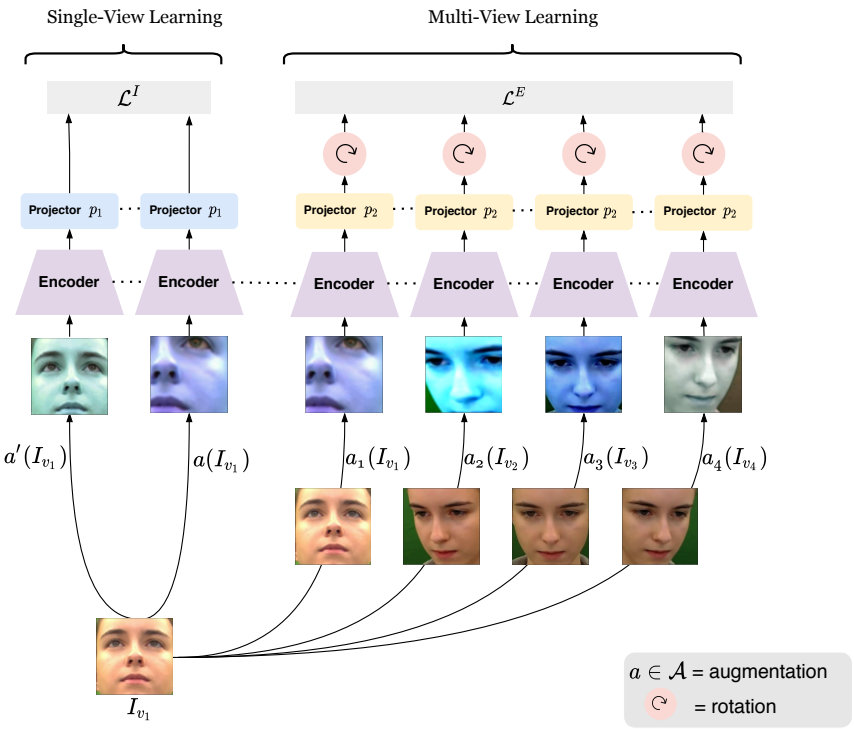


Figure 2: **Method schematic.** For synchronous view frames $\{I_{v_i}\}_{i=1}^4$, the above figure illustrates invariant and equivariant positive pairs anchored only for view v_1. The left branch shows *single-view* learning ($\mathcal{L}^I$) and right branch illustrates *multi-view* learning using four views ($\mathcal{L}^E$). All images (after augmentation, $a \in \mathcal{A}$) are passed through a shared CNN encoder network, followed by MLP projectors (either p_1 or p_2) depending on the type of input positive pair. The embeddings for multi-view learning are further multiplied by an appropriate rotation matrix. More details in Section 2.1.

where, $z_{v_i}^b = p_1(E(I_{v_i}^b))$, $z_{v_i}^{\prime b} = p_1(E(I_{v_i}^{\prime b}))$, $\texttt{sim}(r,s) = \texttt{exp}\left(\frac{1}{\tau}\frac{r^T s}{||r|| \cdot ||s||}\right)$, $1_{[l \neq b]}$ is an indicator function and τ is the temperature coefficient parameter. It is worth noting that to minimize the loss in Eq. 1, it must hold that $z_{v_i}^b$ and $z_{v_i}^{\prime b}$ needs to be closer, which aligns with our goal of learning invariance to appearance transformations. One challenge, however, is the risk of collapse (e.g., the network could simply learn each person's identity). To avoid this, we create mini-batches such that all samples in a batch are taken from a single participant.

Multi-View Learning. We encourage equivariance in the gaze representations to different camera viewpoints through multi-view learning. To do so, we transform embeddings to a common reference system, chosen as the *screen reference system* used during the EVE data collection. Let $\{R_{C_{v_i}}^S\}$ be the rotation matrix relating the camera viewpoint v_i with the screen reference system.

For each sample $I_{v_i}^b$ in a batch of size B, the positive pair is given as $(I_{v_i}^b, I_{v_j}^b)$ for two distinct camera viewpoints $(v_i, v_j)_{i \neq j}$. All images for viewpoints v_i and v_j are first augmented then passed through encoder E and the projector head $p_2(\cdot)$ which gives embeddings $\hat{z}_{v_i}^b, \hat{z}_{v_j}^b \in R^{3 \times d'}$. These embeddings are further multiplied by corresponding rotation matrices $R_{C_{v_i}}^S$ to project embeddings in the common (screen) reference system. We denote embeddings after rotation as $\{\bar{z}_{v_i}^b, \bar{z}_{v_j}^b\}_{b=1}^B$ such that $\bar{z}_{v_i}^b = R_{C_{v_i}}^S \hat{z}_{v_i}^b$. Therefore, for a batch of size B, our equivariant loss $\mathcal{L}^E$ associated with the positive pair $(\bar{z}_{v_i}^b, \bar{z}_{v_j}^b)$ is as follows:

$$\mathcal{L}^E(\bar{z}_{v_i}^b, \bar{z}_{v_j}^b) = -\texttt{log}\frac{\texttt{sim}(\bar{z}_{v_i}^b, \bar{z}_{v_j}^b)}{\sum_{l=1}^B 1_{[l \neq b]}\,\texttt{sim}(\bar{z}_{v_i}^b, \bar{z}_{v_i}^l) + \sum_{l=1}^B \texttt{sim}(\bar{z}_{v_i}^b, \bar{z}_{v_j}^l)} \tag{2}$$

Overall loss function. Given $|V|$ camera viewpoints, we apply both $\mathcal{L}^I$ and $\mathcal{L}^E$ loss functions to each view. Thus, our overall objective function for a batch size of B becomes:

$$\mathcal{L}^O = \frac{1}{2B}\sum_{i=1}^{|V|}\sum_{b=1}^{B}\left(\mathcal{L}^I(z^b_{v_i}, z'^b_{v_i}) + \mathcal{L}^I(z'^b_{v_i}, z^b_{v_i}) + \sum_{j=1, j\neq i}^{|V|}\mathcal{L}^E(\bar{z}^b_{v_i}, \bar{z}^b_{v_j})\right) \tag{3}$$

2.2 Stage-II: Learning For Gaze Estimation

After pre-training, the encoder learned by *GazeCLR* framework is used for the task of gaze estimation and fine-tuned on a small labeled dataset. To this end, we remove both projection heads p_1 and p_2, and replace them with MLP regressor layers to predict 3D gaze direction. For training MLP regressor, we use the supervised loss function given as

$$\mathcal{L}^{ang} = \frac{180}{\pi}\arccos\left(\frac{\boldsymbol{g}\cdot\hat{\boldsymbol{g}}}{||\boldsymbol{g}||\cdot||\hat{\boldsymbol{g}}||}\right) \tag{4}$$

where $\boldsymbol{g}$ and $\hat{\boldsymbol{g}}$ are the ground-truth and predicted gaze directions, respectively.

3 Experiments

We start by detailing the experimental setup (Sec. 3.1) followed by a brief explanation of considered baselines (Sec. 3.2). Next, we evaluate the performance of our pre-trained encoder and show that representations from *GazeCLR* can help train an accurate gaze estimation model even with a relatively lesser amount of annotations. For this task, we consider the within-dataset setting (Sec. 3.3). We assess the transferable capability of our representations by evaluating them on different domains in linear layer training (frozen encoder) setting, where we considered only a few calibration samples from the test subject, as detailed in Sec. 3.4. Thereafter, we compare *GazeCLR* with existing supervised [15] and unsupervised [16, 17] pre-training methods in Sec. 3.5. Lastly, we probe the semantics of learned *GazeCLR* representations using a well-known t-SNE visualization technique (in Sec. 3.6).

3.1 Setup

We train our *GazeCLR* framework on the EVE [12] dataset, which has videos collected in a constrained indoor setting with four different synchronized and calibrated camera views. It has approximately 12 million frames collected from 54 participants with natural eye movements. Following the splits considered by Park et al. [12], there are 40 subjects in training and 6 subjects in the validation set. We discard the data of test subjects due to the non-availability of labels. We use training subjects for the pre-training stage, *without* using any gaze annotations. For the gaze estimation stage, we evaluate on the data of validation subjects to report the performance. We use all four camera views (i.e., $|V| = 4$) as well as the information about the relative pose between camera and screen (R^S_C) provided with the EVE dataset. Note that our framework can be extended to more number of camera views ($|V| > 4$) using ETH-XGaze [19] dataset. In this paper, we consider pre-training only on EVE dataset as more views add on increased computational demand.

All experiments use ResNet-18 [20] as the encoder network and take the output from the average pooling layer. The encoder is trained from scratch. Following Chen et al. [6], both projection heads $p_1(\cdot)$ and $p_2(\cdot)$ are two-layer MLP networks with ReLU non-linearity. The output dimensions for the first and second layers are 512 and 180, respectively. The input image size is 128×128.

Training details. *GazeCLR* is trained using SGD optimizer with initial learning rate $= 0.03$, momentum $= 0.9$, and cosine annealing [21] for the learning rate decay. We use a single 1080 GeForce GTX GPU for training, with a batch size of 128, and train for 50K iterations. Our mini-batch is made up of samples from a single participant. The temperature coefficient τ is set to 0.1. For the augmentation transformations $\mathcal{A}$, we apply random spatial cropping and resizing, gaussian blur, color perturbation ($p = 0.8$) on brightness, contrast, saturation and hue, grayscale conversion ($p = 0.2$) and auto-contrast ($p = 0.5$).

Data pre-processing. We use face images available in the EVE dataset, obtained after applying a data-normalization procedure [22, 23]. The normalization pipeline transforms the gaze annotation to a normalized camera space through a rotation matrix M. Note that we post-multiply R_C^S with M^{-1} as R_C^S is defined w.r.t. the original camera reference frame, i.e., $\bar{z}_v = R_{C_v}^S (M)^{-1} \hat{z}_v$.

We train *GazeCLR* framework in two different settings: (i) *GazeCLR (Equiv)*: where we only consider equivariance through the loss function $\mathcal{L}^E$ and (ii) *GazeCLR (Inv+Equiv)*: where we considered both invariance and equivariance with equal weights using the overall objective $\mathcal{L}^O$. We present the performance of both training setups in all the considered experimental settings. Observe that, *GazeCLR (Inv)* trained with only $\mathcal{L}^I$ loss function is equivalent to SimCLR [6] baseline method.

3.2 Baselines

We compare our approach with six following baselines: (i) *w/o Pre-training*, i.e., an encoder is initialized using random weights, (ii) the vanilla *Autoencoder*, which has an encoder network that consists of same encoder layers as *GazeCLR* and five DenseNet [24] deconvolution blocks as decoder, and is trained with L2 loss, (iii) *Novel View Synthesis* [25] framework is trained on our dataset using same architecture as the auto-encoder, (iv) BYOL [7], (v) SimCLR [6] and (vi) *Fully-Supervised* is a ResNet-18 model trained on the whole EVE training data and represents possibly an upper bound for the performance of *GazeCLR*. For SimCLR and BYOL, we use the same augmentation set as in our proposed method.

3.3 Within-dataset Evaluation

For within-dataset evaluation, we perform pre-training on the training split of the EVE dataset without using labels. Then we adapt the pre-trained encoder for the gaze estimation on a small subset of labeled data. Precisely, we took five training subjects out of 40 (which form around 10% samples out of the whole EVE dataset) for the supervised gaze estimation stage and called it "MiniEVE". We validate on a fixed subject data chosen from training subjects and report the final performance for validation subjects.

Table 1 shows the mean angular errors (in degrees) obtained for different pre-training baselines and the proposed *GazeCLR* method. To this end, we freeze the pre-trained encoder and simply train an MLP regressor using the "MiniEVE" dataset. Note that, for two baselines, Autoencoder and BYOL, we fine-tune the whole end-to-end framework along with the encoder as otherwise, they fail to converge when only their representations are used. We indicate this behavior in Table 1, using the *Frozen* column as ✓ if encoder is frozen otherwise as ✗.

We observe that our method *GazeCLR* outperforms other pre-training baseline methods by only training an MLP regressor on the small amount of labeled data ("MiniEVE" is $\sim 10\%$ of whole data). Specifically, it can be seen that the performance achieved from *GazeCLR* helps in closing the gap with the fully-supervised baseline. Our method *GazeCLR (Inv+Equiv)* shows a relative improvement of 25.1% compared to popular contrastive learning method SimCLR [6]. Additionally, *GazeCLR (Equiv)* shows a boost of 26.4% relative improvement over the SimCLR approach, suggesting that equivariant representations are very effective for the gaze estimation task. We hypothesize that since we utilize similar augmentation strategies for creating both single-view and multi-view positive pairs, *GazeCLR (Equiv)* performs almost comparable to *GazeCLR (Inv+Equiv)*.

3.4 Transfer Learning/Cross-dataset Evaluation

We perform a cross-dataset evaluation using a few-shot personalized gaze estimation to further demonstrate the cross-data generalization capabilities of the learned representations. We evaluate *GazeCLR* representations on two domain datasets different from pre-training data: MPIIGaze [1] and Columbia [14]. **MPIIGaze** is a challenging dataset that has higher inter-subject variations. We use the standard evaluation subset MPIIFaceGaze [2] containing around 37667 images captured from 15 subjects. The **Columbia** dataset consists of 5880 images collected from 56 subjects and is known to have high head pose variations.

To measure the quality of learned representations, we use *Linear Layer Training (LLT)* protocol, in which we freeze the trained encoder and learn a linear regressor on the target dataset. For this

Table 1: **Within-dataset Evaluation.** We report the mean angular errors (MAE) in degrees for within-dataset evaluation for the gaze estimation task. The "EVE" shows the whole EVE data while "MiniEVE" indicates a small subset data. The Frozen column is ✓ if pre-trained encoder is frozen, otherwise fine-tuned ✗. The best performing method is shown in **bold** and second best is underlined.

Method	Pre-Train Data	Task Data	Frozen	MAE ↓ (degrees)
w/o Pre-training	EVE	MiniEVE	✗	8.47
Autoencoder	EVE	MiniEVE	✗	6.91
Novel View Synthesis [25]	EVE	MiniEVE	✓	6.79
BYOL [7]	EVE	MiniEVE	✗	8.35
SIMCLR [6]	EVE	MiniEVE	✓	6.57
GazeCLR (Equiv)	EVE	MiniEVE	✓	**4.83**
GazeCLR (Inv+Equiv)	EVE	MiniEVE	✓	<u>4.92</u>
Fully-Supervised	-	EVE	✗	4.15

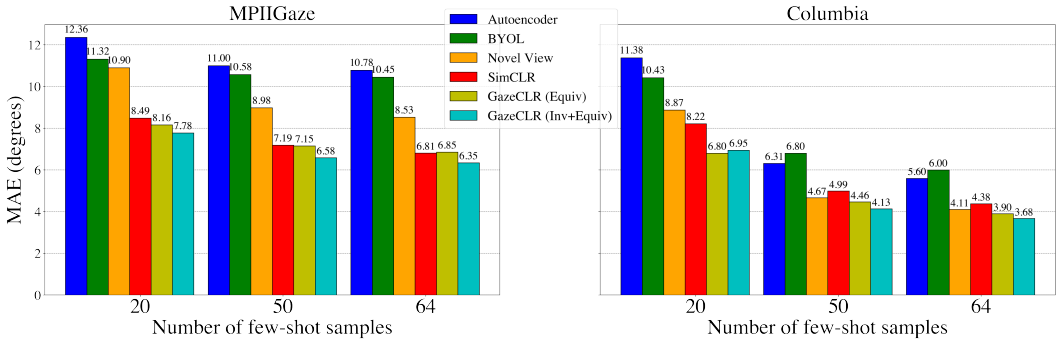


Figure 3: **Transfer Learning Evaluation.** Performance evaluation using *Linear Layer Training (LLT)* protocol for both MPIIGaze and Columbia dataset under different few-shot settings. Each bar is computed by averaging over 10 runs. Best viewed in color and after zooming.

experiment, we investigate under a few-shot setting where we sample a few calibration samples from the test subject for adaptation, and evaluate on the remaining samples of the same test subject.

Figure 3 shows the mean angular errors for LLT protocol on 20-shot, 50-shot, and 64-shot gaze estimation. We first extract the gaze representations of a few calibration samples for each subject and learn a linear model on top of these representations. We evaluate the trained model on the remaining samples of the subject. We repeat above 10 times for each subject on both datasets and report mean angular errors for the same in Figure 3.

Observe that both proposed *GazeCLR* variants outperform all other baselines in all few-shot settings for both datasets. Moreover, *GazeCLR(Equiv)* gives an relative improvement of around 17.2% over SimCLR with only 20 calibration samples for Columbia. We hypothesize that this behavior is due to high head-pose variations within Columbia and it suggests that: a) learning equivariance over multi-views is beneficial for the GazeCLR framework in improving performance, and b) *GazeCLR* representations are relatively more generalizable for cross-domain datasets than other baselines.

3.5 Comparison with state-of-the-art gaze representation learning

We further compare *GazeCLR* with existing state-of-the-art unsupervised [16, 17] and supervised [15] gaze representation learning methods. For a fair comparison, we adopt the same evaluation protocols as used by these baseline methods and compare the *GazeCLR* performance against their performance.

***GazeCLR* vs Unsupervised Pre-training [16, 17].** We follow the same evaluation protocol as [16]. 5-fold and leave-one-out (15-fold) evaluations are used for the Columbia and MPIIGaze datasets, respectively. In each fold, we freeze *GazeCLR* encoder and extract representations for randomly selected 50 samples with annotations and learn a simple MLP-based gaze estimator on top of that. We repeat performance evaluation for 10 times and report mean angular errors in Table 2. Note that

Table 2: ***GazeCLR* vs [16, 17]:** Comparison of *GazeCLR* with other unsupervised gaze representation learning methods [16, 17] for 50-shot gaze estimation. † denotes the method that uses additional head pose information. The metric reported is mean angular errors averaged over 10 runs (in degrees).

Method	Pre-Train Data	MPIIGaze	Columbia
Yu et al. [16]†	Columbia	-	8.9
Sun et al. [17]	MPIIGaze	8.5	-
Sun et al. [17]	Columbia	-	7.0
***GazeCLR* (Equiv)**	EVE	7.0	**6.1**
***GazeCLR* (Inv+Equiv)**	EVE	**6.5**	6.6

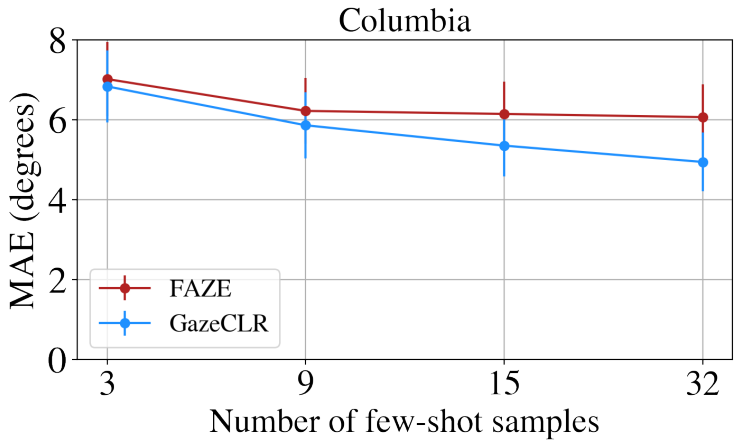


Figure 4: ***GazeCLR* vs FAZE [15].** Comparison of *GazeCLR* with supervised pre-training baseline (FAZE) for various few-shot settings, on the Columbia dataset. The plot shows mean angular error (MAE, in degrees) and standard error bars *versus* number of few-shot samples, reported after 10 runs.

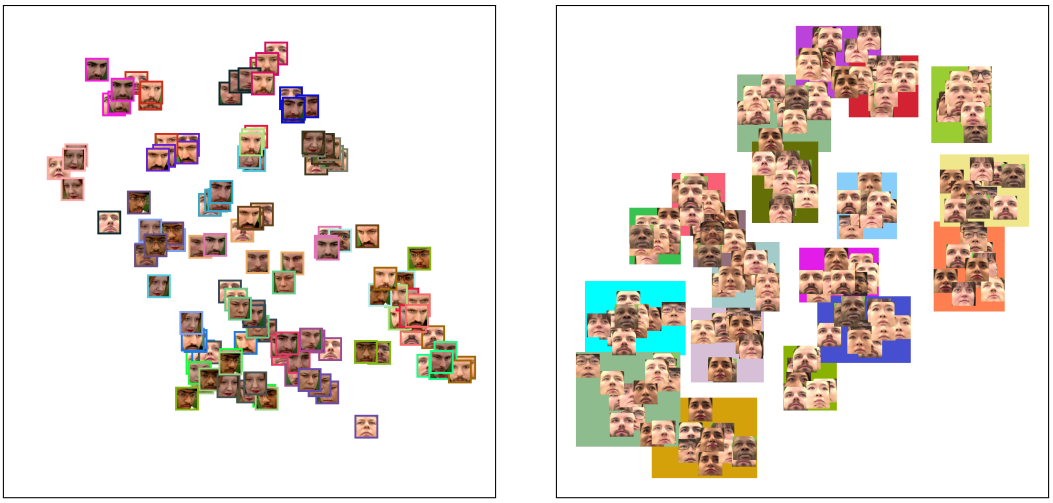
previous methods [16, 17] exploit left and right eye patches to get SSL signal, whereas our approach relies on face patches obtained from multiple camera viewpoints.

In Table 2, we compare against the best-performing models of Yu et al. [16] and Sun et al. [17], for the 50-shot gaze estimation. Notice that our method outperforms baselines with absolute improvements of 2° and 0.9° on MPIIGaze and Columbia, respectively. It is worth emphasizing that our method is pre-trained on a different dataset than both evaluation datasets, unlike baseline approaches. Again, it illustrates the strength of our approach in learning semantically meaningful representations for generalizable to other domains. Moreover, note that Yu et al. [16] use additional head-pose information, unlike our method.

***GazeCLR* vs Supervised Pre-training [15].** We evaluate the effectiveness of *GazeCLR* representations using the MAML framework [26], similar to FAZE [15]. For both *GazeCLR* and FAZE, we train a MAML-based gaze estimator on the representations for subjects from the GazeCapture [3] dataset. Then, we adapt the gaze estimator model to each test subject of Columbia with k calibration samples and test on the remaining samples. Figure 4 depicts the performance comparison of *GazeCLR* with FAZE [15] for four different values of k. It can be seen that our method consistently outperforms supervised pre-training baseline FAZE, for all values of k. Notably, our framework uses *zero* labeled information to obtain gaze representations, unlike FAZE, which is pre-trained using $\sim 2M$ labeled samples from the GazeCapture dataset.

3.6 Visualization of Gaze Representations

To further investigate the quality of learned representations, we project the gaze representations into 2-dimensions using t-SNE [27] algorithm as shown in Figure 5. In Fig 5(a), we compute 2D visualization of equivariant representations obtained after applying rotation matrices, i.e., $\bar{z}$.



(a) Representations after applying rotation matrices, i.e., $\bar{z}$

(b) Representation obtained from the output of encoder, $z = \mathrm{E}(\cdot)$

Figure 5: **t-SNE visualization.** Qualitative visualization of gaze representations in 2-dimensional space using the t-SNE algorithm. (a) shows the visualization of projection embeddings for multi-view images obtained after applying rotation matrices, i.e., $\bar{z}$. The images with the same timestamp for all four views are highlighted using the same border color. (b) depicts representations for the output of encoder network, i.e., $z = \mathrm{E}(\cdot)$ obtained for images from single camera viewpoint. Best viewed in color and after zooming.

Projections in Fig 5(a) clearly demonstrate that gaze direction is invariant to the viewpoint, as images at the same timestamp from different views are mapped closer (shown with the same color border). In Fig 5(b), we apply t-SNE algorithm on gaze representations obtained at output of encoder network, i.e., $z = \mathrm{E}(\cdot)$, for images from single camera viewpoint. Projections corresponding to roughly similar gaze directions are naturally clustered and highlighted with different background colors. Also, we observe clear patterns in the learned feature space where images within close vicinity are invariant to the subject's identity, showing invariance towards appearances.

4 Related Work

Gaze Estimation. Gaze estimation methods are built using large-scale datasets either having 2D target labels [3, 28] or 3D gaze directions [29, 30, 1]. Broadly, gaze estimation methods can be divided into two categories: appearance methods [31], which directly map image pixels to 3D gaze direction, and model methods which rely on eye-geometry [32]. Appearance methods perform better than traditional model methods in real-world settings [32, 1].

Recent progress in appearance methods relies heavily on deep learning to map eye/face images to gaze directions [1, 2]. Furthermore, a few gaze methods are hybrid. For instance, Park et al. [33] extracted the relevant eye landmarks from the images and then used these features to train model gaze estimators. Other than eye images, several works [3, 34, 35, 36] exploit both eyes and face images in computing gaze direction. Nevertheless, both appearance methods and hybrid extensions require a huge amount of labeled data to achieve their potential in terms of accuracy.

As a result of huge label dependence by appearance methods, efforts have been made in the direction of few-shot gaze estimation. In particular, Liu et al. [37] exploit a two-branch network to predict differential gaze between two images and use few calibration samples during inference. Furthermore, [15, 38] disentangle gaze from other nuisance factors via training an encoder-decoder architecture [39] to learn gaze specific representations. Recent approaches [40, 41, 42] leverage labeled source domain and unsupervised domain adaptation for improving the performance of gaze estimation task on target domain. In a similar spirit, our work attempts to reduce the amount of required label information via learning gaze representations without relying on gaze labels and utilizing multi-view data [12, 19].

Self-Supervised Learning. The goal of self-supervised representation learning is to learn good visual representations from a large collection of unlabeled images. Earlier works in SSL [43, 44, 45, 46] used pretext tasks to learn generalizable semantic representations. Some of the recent works [47, 5, 6, 48, 49, 50, 7] have shown a great success on several vision tasks, e.g, image classification [8, 51], object detection [9], semantic segmentation [10], and pose estimation [25]. The recent work by Spurr et al. [18] extends SSL to hand pose estimation through geometric equivariance representations. Tian et al. [52] propose to use more than two views to learn invariant representations through contrastive learning.

Recently, a few unsupervised methods have been proposed to learn gaze-specific representations. Specifically, Yu et al. [16] exploit the gaze redirection task to train a gaze estimation model using paired eye images of the same subject. Similarly, Sun et al. [17] proposed a cross-encoder method to utilize patches of left and right eye images of the same subject as the self-supervised signal. Gideon et al. [53] is an extended version of [17] utilizing multi-view images and learning features representing head pose and relative gaze to improve in-domain few-shot gaze estimation performance. However, unlike our work, these methods employ an encoder-decoder framework and thus require a relatively large number of parameters. Also, contrastive SSL approaches are computationally efficient compared to generative SSL approaches [54].

5 Conclusion

We presented *GazeCLR*, a contrastive learning framework for gaze representations using multi-view camera images. Our framework induces invariance and equivariance properties simultaneously in the learned representations and is effective for gaze estimation task in various settings. Furthermore, we showed that *GazeCLR* representations have the potential to be effective across different domain datasets using only a few calibration samples. *GazeCLR* is a general framework for equivariant representation learning, thus can be explored in the future for other geometry-based applications such as head pose estimation.

Acknowledgments and Disclosure of Funding

This research was supported by the National Eye Institute of the National Institutes of Health (NIH) under award number R01EY030952-01A1. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. We thank Mohit Yadav for helpful discussions.

References

[1] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "Appearance-based gaze estimation in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4511–4520, 2015.

[2] X. Zhang, Y. Sugano, M. Fritz, and A. Bulling, "It's written all over your face: Full-face appearance-based gaze estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 51–60, 2017.

[3] K. Krafka, A. Khosla, P. Kellnhofer, H. Kannan, S. Bhandarkar, W. Matusik, and A. Torralba, "Eye tracking for everyone," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2176–2184, 2016.

[4] S. Ghosh, A. Dhall, M. Hayat, J. Knibbe, and Q. Ji, "Automatic gaze analysis: A survey of deep learning based approaches," *arXiv preprint arXiv:2108.05479*, 2021.

[5] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

[6] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, pp. 1597–1607, PMLR, 2020.

[7] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in Neural Information Processing Systems*, vol. 33, pp. 21271–21284, 2020.

[8] M. Caron, P. Bojanowski, A. Joulin, and M. Douze, "Deep clustering for unsupervised learning of visual features," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 132–149, 2018.

[9] E. Crawford and J. Pineau, "Spatially invariant unsupervised object detection with convolutional neural networks," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 3412–3420, 2019.

[10] T. Moriya, H. R. Roth, S. Nakamura, H. Oda, K. Nagara, M. Oda, and K. Mori, "Unsupervised segmentation of 3d medical images based on clustering and deep representation learning," in *Medical Imaging 2018: Biomedical Applications in Molecular, Structural, and Functional Imaging*, vol. 10578, p. 1057820, International Society for Optics and Photonics, 2018.

[11] S. Chopra, R. Hadsell, and Y. LeCun, "Learning a similarity metric discriminatively, with application to face verification," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 539–546, IEEE, 2005.

[12] S. Park, E. Aksan, X. Zhang, and O. Hilliges, "Towards end-to-end video-based eye-tracking," in *European conference on computer vision*, pp. 747–763, Springer, 2020.

[13] Tobii Pro AB, "Tobii pro lab." Computer software, 2014.

[14] B. A. Smith, Q. Yin, S. K. Feiner, and S. K. Nayar, "Gaze locking: passive eye contact detection for human-object interaction," in *Proceedings of the 26th annual ACM symposium on User interface software and technology*, pp. 271–280, 2013.

[15] S. Park, S. D. Mello, P. Molchanov, U. Iqbal, O. Hilliges, and J. Kautz, "Few-shot adaptive gaze estimation," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 9368–9377, 2019.

[16] Y. Yu and J.-M. Odobez, "Unsupervised representation learning for gaze estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7314–7324, 2020.

[17] Y. Sun, J. Zeng, S. Shan, and X. Chen, "Cross-encoder for unsupervised gaze representation learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3702–3711, 2021.

[18] A. Spurr, A. Dahiya, X. Wang, X. Zhang, and O. Hilliges, "Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11230–11239, 2021.

[19] 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 *European Conference on Computer Vision*, pp. 365–381, Springer, 2020.

[20] 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*, pp. 770–778, 2016.

[21] I. Loshchilov and F. Hutter, "Sgdr: Stochastic gradient descent with warm restarts," *arXiv preprint arXiv:1608.03983*, 2016.

[22] Y. Sugano, Y. Matsushita, and Y. Sato, "Learning-by-synthesis for appearance-based 3d gaze estimation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1821–1828, 2014.

[23] X. Zhang, Y. Sugano, and A. Bulling, "Revisiting data normalization for appearance-based gaze estimation," in *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, pp. 1–9, 2018.

[24] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700–4708, 2017.

[25] H. Rhodin, M. Salzmann, and P. Fua, "Unsupervised geometry-aware representation for 3d human pose estimation," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 750–767, 2018.

[26] C. Finn, P. Abbeel, and S. Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *International Conference on Machine Learning*, pp. 1126–1135, PMLR, 2017.

[27] L. Van der Maaten and G. Hinton, "Visualizing data using t-sne.," *Journal of machine learning research*, vol. 9, no. 11, 2008.

[28] Q. Huang, A. Veeraraghavan, and A. Sabharwal, "Tabletgaze: dataset and analysis for unconstrained appearance-based gaze estimation in mobile tablets," *Machine Vision and Applications*, vol. 28, no. 5, pp. 445–461, 2017.

[29] 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)*, pp. 334–352, 2018.

[30] 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*, pp. 255–258, 2014.

[31] K.-H. Tan, D. J. Kriegman, and N. Ahuja, "Appearance-based eye gaze estimation," in *Sixth IEEE Workshop on Applications of Computer Vision, 2002.(WACV 2002). Proceedings.*, pp. 191–195, IEEE, 2002.

[32] D. W. Hansen and Q. Ji, "In the eye of the beholder: A survey of models for eyes and gaze," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 3, pp. 478–500, 2009.

[33] S. Park, X. Zhang, A. Bulling, and O. Hilliges, "Learning to find eye region landmarks for remote gaze estimation in unconstrained settings," in *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, pp. 1–10, 2018.

[34] 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.

[35] T. Fischer, H. J. Chang, and Y. Demiris, "RT-GENE: Real-Time Eye Gaze Estimation in Natural Environments," in *European Conference on Computer Vision*, pp. 339–357, September 2018.

[36] Y. Cheng, S. Huang, F. Wang, C. Qian, and F. Lu, "A coarse-to-fine adaptive network for appearance-based gaze estimation.," in *AAAI*, pp. 10623–10630, 2020.

[37] G. Liu, Y. Yu, K. A. F. Mora, and J.-M. Odobez, "A differential approach for gaze estimation with calibration.," in *BMVC*, vol. 2, p. 6, 2018.

[38] 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. 13127–13138, 2020.

[39] G. E. Hinton, A. Krizhevsky, and S. D. Wang, "Transforming auto-encoders," in *International conference on artificial neural networks*, pp. 44–51, Springer, 2011.

[40] Y. Wang, Y. Jiang, J. Li, B. Ni, W. Dai, C. Li, H. Xiong, and T. Li, "Contrastive regression for domain adaptation on gaze estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19376–19385, 2022.

[41] Y. Bao, Y. Liu, H. Wang, and F. Lu, "Generalizing gaze estimation with rotation consistency," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4207–4216, 2022.

[42] S. Jindal and X. E. Wang, "Cuda-ghr: Controllable unsupervised domain adaptation for gaze and head redirection," *arXiv preprint arXiv:2106.10852*, 2021.

[43] R. Zhang, P. Isola, and A. A. Efros, "Colorful image colorization," in *European conference on computer vision*, pp. 649–666, Springer, 2016.

[44] M. Noroozi and P. Favaro, "Unsupervised learning of visual representations by solving jigsaw puzzles," in *European conference on computer vision*, pp. 69–84, Springer, 2016.

[45] M. Noroozi, H. Pirsiavash, and P. Favaro, "Representation learning by learning to count," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 5898–5906, 2017.

[46] C. Doersch, A. Gupta, and A. A. Efros, "Unsupervised visual representation learning by context prediction," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[47] I. Misra and L. v. d. Maaten, "Self-supervised learning of pretext-invariant representations," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6707–6717, 2020.

[48] X. Chen, H. Fan, R. Girshick, and K. He, "Improved baselines with momentum contrastive learning," *arXiv preprint arXiv:2003.04297*, 2020.

[49] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in Neural Information Processing Systems*, vol. 33, pp. 9912–9924, 2020.

[50] X. Chen and K. He, "Exploring simple siamese representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15750–15758, 2021.

[51] R. Dangovski, L. Jing, C. Loh, S. Han, A. Srivastava, B. Cheung, P. Agrawal, and M. Soljacic, "Equivariant self-supervised learning: Encouraging equivariance in representations," in *International Conference on Learning Representations*, 2022.

[52] Y. Tian, D. Krishnan, and P. Isola, "Contrastive multiview coding," in *European conference on computer vision*, pp. 776–794, Springer, 2020.

[53] J. Gideon, S. Su, and S. Stent, "Unsupervised multi-view gaze representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5001–5009, 2022.

[54] X. Liu, F. Zhang, Z. Hou, L. Mian, Z. Wang, J. Zhang, and J. Tang, "Self-supervised learning: Generative or contrastive," *IEEE Transactions on Knowledge and Data Engineering*, 2021.