

# Towards Self-Supervised Gaze Estimation

Arya Farkhondeh<sup>1,3</sup>

farkhondeh.1860768@studenti.uniroma1.it

Cristina Palmero<sup>2,3</sup>

cpalmec7@alumnes.ub.edu

Simone Scardapane<sup>1</sup>

simone.scardapane@uniroma1.it

Sergio Escalera<sup>2,3</sup>

sergio@maia.ub.es

<sup>1</sup> Sapienza University of Rome  
Rome, Italy

<sup>2</sup> University of Barcelona  
Barcelona, Spain

<sup>3</sup> Computer Vision Center (CVC)  
Barcelona, Spain

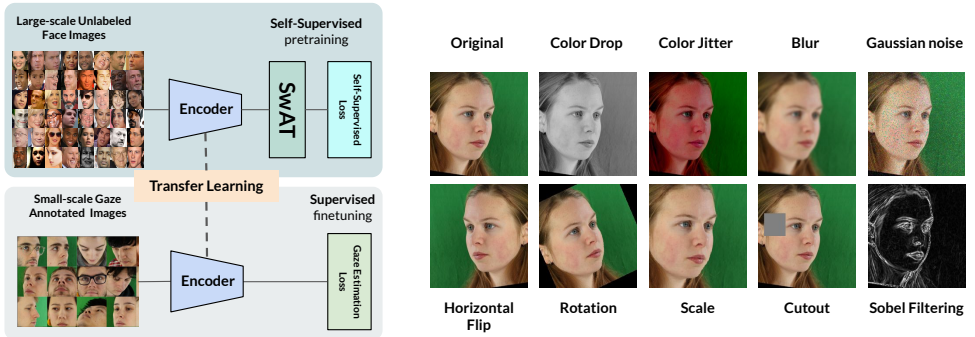
## Abstract

Recent joint embedding-based self-supervised methods have surpassed standard supervised approaches on various image recognition tasks such as image classification. These self-supervised methods aim at maximizing agreement between features extracted from two differently transformed views of the same image, which results in learning an invariant representation with respect to appearance and geometric image transformations. However, the effectiveness of these approaches remains unclear in the context of gaze estimation, a structured regression task that requires equivariance under geometric transformations (e.g., rotations, horizontal flip). In this work, we propose SwAT, an equivariant version of the online clustering-based self-supervised approach SwAV, to learn more informative representations for gaze estimation. We demonstrate that SwAT, with ResNet-50 and supported with uncurated unlabeled face images, outperforms state-of-the-art gaze estimation methods and supervised baselines in various experiments. In particular, we achieve up to 57% and 25% improvements in cross-dataset and within-dataset evaluation tasks on existing benchmarks (ETH-XGaze, Gaze360, and MPIIFaceGaze).

## Introduction

Appearance-based gaze estimation remains a non-trivial problem to solve within the computer vision field due to the large variability across appearance and geometric factors. Convolutional neural network (CNN) based methods [6, 8, 19, 22, 28, 29, 41] have achieved promising performances fueled by large-scale datasets [19, 22, 45]. Nonetheless, there is still a large gap to achieve a desirable performance especially when it comes to generalizing to unseen distributions with novel head poses, appearances, geometry, and illuminations. One way to address this problem is through the acquisition of even larger in-the-wild, gaze-annotated datasets with more variability. However, collecting data with accurate gaze annotations is an unscalable and laborious process that requires controlled conditions, complicated setups, tedious camera calibration, and subject recruitment. An inexpensive solution is therefore needed to extend variability in terms of appearance and geometric factors.

Recently, joint embedding-based self-supervised methods, including contrastive and non-contrastive, have obtained remarkable accuracy on various vision tasks, such as image classification [9, 8, 4, 5, 15], object detection [36], and hand-pose estimation [52]. These approaches have proven successful at learning generalizable features by leveraging large-scale



**Figure 1: Left. Global view of our approach.** In the first stage, we pretrain an encoder via an online clustering approach on a large-scale set of unlabeled face images while encouraging equivariance through our proposed method (SwAT). In the second stage, we transfer the learned knowledge from the first stage and fine-tune on a small-scale set of gaze-annotated images. **Right. Transformation Catalog.** The appearance and geometric transformations explored in this work for self-supervised representation learning.

unlabeled data [18, 23]. Similarly, these methods could leverage the vast amount of unlabeled face images that are publicly available on the Internet to learn useful representations for appearance-based gaze estimation. However, little attention has been paid to investigating their effectiveness for the gaze estimation task. Therefore, the main goal of this work is to explore the efficacy of a self-supervised approach in the context of gaze estimation to reduce the reliance on large-scale gaze-annotated data that is laborious to acquire. We specifically focus on full-face instead of eye-only images as input since the face provides auxiliary information [22, 28, 32].

In a nutshell, self-supervised learning aims at solving a pretext task to learn a useful representation. The representation is then used in downstream tasks via transfer learning. The common pretext task among (non-)contrastive self-supervised methods (e.g., SimCLR [4], MoCo [17], SwAV [3], BYOL [15], and VICReg [2]) is to enforce consistency between features extracted from two differently transformed views of the same image. As a result, the feature extractor is encouraged to learn an invariant representation with respect to the image-space transformations, such as appearance (e.g., color jitter) and geometric (e.g., horizontal flip). Although invariance might be a desired property for most image recognition tasks, the structured regression task of gaze estimation requires equivariance under geometric transformations. In fact, applying geometric transformations to a face/eye image results in respective changes in gaze direction. Thus, in this work, our goal is to learn an equivariant representation under geometric transformations to align with the gaze estimation task.

In this paper, we propose **Swapping Affine Transformations (SwAT)**, a novel method to achieve the desired property of equivariance. It can be thought of as a plug-and-play method that can be added to any joint embedding-based self-supervised approach. As Fig. 1 depicts, we perform self-supervised pretraining on large-scale unlabeled face images while encouraging equivariance through SwAT. Then, we transfer the learned knowledge to the downstream gaze estimation task and finetune with gaze labels. Intuitively, SwAT allows the feature extractor to transfer the image-space geometric transformation to the representation output which preserves the intrinsic structure of the transformations.

Our proposed self-supervised approach potentially deconcentrates research in gaze esti-

mation from the non-trivial process of large-scale annotated data collection towards effectively leveraging widely available large-scale unlabeled data. More importantly, leveraging such unlabeled data with more variety enhances the generalizability of gaze estimation models upon novel distributions. We show that the equivariance property provided by SwAT leads to learning better representations for gaze estimation, compared to other pretraining regimes. We also show that the unsupervised features provided by SwAT surpass the commonly used ImageNet supervised features in gaze estimation. We perform extensive experiments to verify the effectiveness of our approach under various challenging evaluation settings. We demonstrate that SwAT outperforms the supervised baselines in low-data regimes where only a few annotations (10% and 30%) are available. Supported with unlabeled data, SwAT achieves state-of-the-art results on existing benchmarks and improves the supervised baselines for cross- and within- dataset evaluation tasks by 57% and 25%, respectively.

## 2 Related Work

**Self-Supervised Learning.** Early self-supervised approaches attempted to learn useful representations from unlabeled data via solving handcrafted pretext tasks such as Jigsaw puzzle [27], colorization [40], transformation prediction [0, 24], and inpainting [53]. More recently, contrastive-based methods [0, 17, 25] have achieved notable results on various computer vision tasks such as image classification. However, these methods are inherently computationally inefficient as they require pairwise contrasts with a large set of negative examples. Consequently, non-contrastive approaches [0, 8, 6, 15] are receiving special attention. Clustering-based approaches such as SwAV [9] discriminate between groups of images with similar features instead of individual images. However, both contrastive and non-contrastive methods are designed to learn invariant representations under image transformations, while gaze estimation requires equivariance under geometric transformations. Hence, in this work, we extend SwAV [9] via introducing equivariance under geometric transformations. Equivariance in self-supervised learning is starting to attract attention [11, 32, 37]. Despite their proven effectiveness, these methods bear some limitations that do not align with our assumptions. While our goal is to promote equivariance for multiple affine transformations, Dangovski et al. [11]’s work is limited to a single transformation and Xie et al. [37]’s method is not scalable as the number of transformations increments. Most similarly, Spurr et al. [32] propose an equivariance formulation for the task of 3D hand-pose estimation. However, their equivariance formulation together with a contrastive loss explicitly pushes apart the pseudo-negative pairs that may include faces with similar affine information, gaze, and head directions.

**Appearance-based Gaze Estimation.** Recent progress in appearance-based gaze estimation has been mainly achieved via collecting large-scale datasets [12, 19, 22, 26, 45], task-specific tailored architectures [6, 7, 11, 29], and data normalization methods [43, 44]. Apart from supervised gaze estimation, weakly-supervised and unsupervised methods have started to receive more attention in gaze estimation. Kothari et al. [21] propose a weakly-supervised approach based on videos of people looking at each other. MTGLS [13] utilizes off-the-shelf models to obtain pseudo labels for unlabeled eye images in order to learn a gaze representation. Recent generative-based unsupervised gaze estimation approaches [33, 39] make use of unlabeled eye images to learn gaze representations. Nevertheless, these approaches have limitations as they require supervision in the form of paired eye images of the same person [33, 39] with similar head-pose [39]. Wu et al. [35] employ self-supervision as an auxiliary task for supervised gaze estimation. Unlike the previous methods, we pretrain a standard CNN architecture for gaze estimation in a self-supervised fashion via leveraging

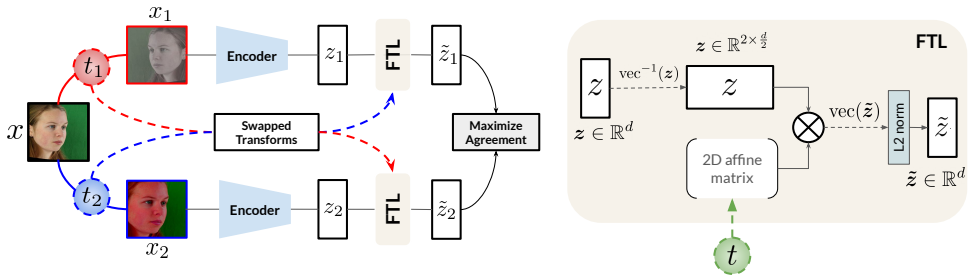


Figure 2: **Left.** Two differently transformed versions of the same image  $x$  are obtained via applying two different sets of transformations i.e.,  $x_1 = t_1(x)$  and  $x_2 = t_2(x)$ . An encoder is used to map the transformed views to vector representations,  $z_1$  and  $z_2$ . To achieve equivariance, we equalize  $z_1$  and  $z_2$  in terms of affine information. To do so, we swap the affine transformations applied in image space  $t_1$  and  $t_2$ , then we use the feature transform layer (FTL) to apply the swapped transformations to the feature vectors i.e.,  $\tilde{z}_1 = \text{FTL}(t_2, z_1)$  and  $\tilde{z}_2 = \text{FTL}(t_1, z_2)$ . Then, we maximize agreement between the resulting feature vectors,  $\tilde{z}_1$  and  $\tilde{z}_2$ . **Right.** Details of the feature transform layer (FTL).  $\text{vec}^{-1}(z)$  transforms  $z$  from 1D to 2D in order to enable matrix-matrix multiplication with the 2D affine matrix, resulting in  $\tilde{z}$ . Then,  $\text{vec}(\tilde{z})$  transforms back  $\tilde{z}$  from 2D to 1D. L2-norm is then applied.

large-scale unlabeled face images. Our approach is less complex while more scalable as it does not make any assumption on the kind of unlabeled data and does not require multiple auxiliary losses for training as in [63, 69]. Furthermore, in contrast to previous unsupervised works that use eye images, we use full-face images, which have been proven to contain useful auxiliary information (e.g., head-pose, geometric features) for gaze estimation [22, 28, 42].

## 3 Method

As Fig. 1 depicts, our goal is to pretrain an encoder on large-scale unlabeled face images using a self-supervised approach (Sec. 3.1) while encouraging equivariance via SwAT (Sec. 3.2). Afterward, we transfer the knowledge to the gaze estimation task via supervised finetuning (Sec. 3.3).

### 3.1 Self-Supervised Pretraining

In this work, as pretext task, we aim at maximizing the mutual information between the features from two different views of the same image. As shown in Fig. 2 (left), two differently transformed views of an image  $x$  are computed via applying two different sets of transformations i.e.,  $x_1 = t_1(x)$  and  $x_2 = t_2(x)$  where,  $t_1 \sim \mathcal{T}$  and  $t_2 \sim \mathcal{T}$  are sampled from the same transformation catalog  $\mathcal{T}$ . An encoder  $f_\phi(\cdot)$  parameterized by  $\phi$  maps the transformed views to vector representations,  $z_1 = f_\phi(x_1)$  and  $z_2 = f_\phi(x_2)$ . The encoder  $f_\phi(\cdot)$  is composed of a backbone (e.g., ResNet) and a projection head (e.g., MLP). Then, we maximize agreement between the feature vectors using an online clustering-based self-supervised approach called SwAV [9]. SwAV enforces agreement using intermediate cluster assignments computed in an online fashion, where the cluster assignments are treated as the targets to predict from feature vectors. To compute the cluster assignments  $c_1$  and  $c_2$ , the vector representations ( $z_1$  and  $z_2$ ) are compared to a set of  $M$  learnable prototype vectors  $P_\psi = \{p_1, \dots, p_M\}$ , parameterized by  $\psi$ . Maximizing agreement is achieved via swapping the computed cluster assignments and predicting them using feature vectors. The idea is to predict the cluster assignment  $c_1$  from the feature  $z_2$ , and  $c_2$  from  $z_1$ . Intuitively, if two feature vectors contain

mutual information then it should be possible to predict the cluster assignment  $c_1$  ( $c_2$ ) from the other feature  $z_2$  ( $z_1$ ). The self-supervised loss function is as follows:

$$\mathcal{L}_{\text{SwAV}} = \ell(z_1, c_2) + \ell(z_2, c_1), \quad (1)$$

where  $\ell(z, c)$  is the cross entropy loss between the cluster assignments and the probability computed by applying softmax to the dot products of  $z_i$  and prototypes ( $P_\psi$ ), as in Eq. 2. The cross entropy loss measures agreement between a feature and cluster assignment.  $\ell(z_i, c_j)$  is defined as follows:

$$\ell(z_i, c_j) = - \sum_m c_j^{(m)} \log \left( \frac{\exp(\frac{1}{\tau} z_i^\top P_m)}{\sum_{m'} \exp(\frac{1}{\tau} z_i^\top P_{m'})} \right), \quad (2)$$

where  $\tau$  is a temperature parameter and  $m$  denotes the  $m$ th prototype. The overall loss function (Eq. 1) is minimized with respect to both parameters of the encoder  $\phi$  and trainable prototypes  $\psi$ . The method is online since only the features within a batch are used to compute the cluster assignments. To avoid trivial solutions i.e., assigning the same cluster for every image within a batch, score adjustment is performed using an optimal transport algorithm, namely Sinkhorn-Knopp [9]. It encourages equipartition guaranteeing that the cluster assignments are distinct for images within a batch.

## 3.2 Equivariant Representation Learning

Similar to other (non-)contrastive self-supervised approaches, the SwAV formulation (Sec. 3.1) encourages invariance under appearance and geometric transformations. In image recognition tasks such as image classification, applying geometric transformations ( $t^g$ ) to an image does not change the label. However, in the gaze estimation task, applying geometric transformations in image space results in respective changes in label space. Thus, instead of learning an *invariant* representation, we aim at learning an *equivariant* representation.

**Definition 1 (Equivariance)** A mapping function  $f_\phi : x \rightarrow z$  is said to be equivariant with respect to image-space transformation  $t_1^g$  when mapping the transformed input image,  $f_\phi(t_1^g(x))$ , produces the same result as transforming the vector representation of the input image, i.e.,  $t_F^g(f_\phi(x))$ :

$$f_\phi(t_1^g(x)) = t_F^g(f_\phi(x)), \quad (3)$$

where transformations  $t_1^g$  and  $t_F^g$  are used to apply the same transformation in different spaces i.e., image space and feature space, respectively. Intuitively, the *equivariance* property enables  $f_\phi$  to learn a direct relationship between image space and feature space, thereby preserving the intrinsic structure of the transformations [94].

**Swapping Affine Transformations.** Eq. 1 enforces consistent mapping between two transformed views via intermediate cluster assignments. Abstractly, it aims to maximize the mutual information between the features from two views. Thus, ideally,

$$f_\phi(t_1^g(x)) = f_\phi(t_2^g(x)). \quad (4)$$

The only way that the above equality is satisfied is through encouraging  $f_\phi$  to be invariant with respect to the applied geometric transformations  $t_1^g$  and  $t_2^g$ . Instead, to let the mapping function  $f_\phi$  be equivariant under affine transformations applied in image space, we propose the **Swapping Affine Transformations (SwAT)** method. SwAT achieves equivariance via equalization of vector representations in terms of applied image-space affine transformations. To achieve that, as in Eq. 5 and Fig. 2, we swap the affine transformations applied in image-space, and then we apply them in feature-space via a feature transform layer ( $\mathbb{F}^{\text{TTL}}$ ), detailed later. Thus,

$$\tilde{z}_1 = \mathbb{F}\mathbb{T}\mathbb{L}(t_2^g, z_1), \quad \tilde{z}_2 = \mathbb{F}\mathbb{T}\mathbb{L}(t_1^g, z_2). \quad (5)$$

Intuitively,  $\tilde{z}_1$  and  $\tilde{z}_2$  contain the same affine transformation information. Thus, enforcing consistency between transformation-equalized vector representations prevents  $f_\phi$  from becoming invariant with respect to transformations. In contrast, since unequalized vector representations  $z_1$  and  $z_2$  contain different transformation information, enforcing consistency would result in invariance as in Eq. 4. The self-supervised loss (Eq. 1) becomes:

$$\mathcal{L}_{\text{SwAT}} = \ell(\tilde{z}_1, \tilde{c}_2) + \ell(\tilde{z}_2, \tilde{c}_1), \quad (6)$$

where,

$$\tilde{c}_2 = \tilde{z}_2 P_\psi, \quad \tilde{z}_2 \in \mathbb{R}^d, \quad P_\psi \in \mathbb{R}^{d \times M}. \quad (7)$$

**Feature Transform Layer.** As Fig. 2 (right) depicts, to be able to apply the feature-space equivalent ( $t_F^g$ ) of the image-space transformation ( $t_I^g$ ), we introduce a non-trainable feature transform layer ( $\mathbb{F}\mathbb{T}\mathbb{L}$ ). This layer takes as input the 2D affine transformation matrix  $T_\theta$  (e.g., 2D rotation matrix with angle  $\theta$ ) and 1D feature vector  $z$ . It first transforms  $z$  from 1D to 2D via an inverse vectorization,  $\text{vec}_{2 \times k}^{-1}(z)$ , where  $k = \frac{d}{2}$  and  $d$  is the dimensionality of the projection head. Afterward, it performs a matrix-matrix multiplication, resulting in  $\tilde{z}$ . Finally,  $\tilde{z}$  is transformed back to a 1D feature vector via a vectorization,  $\text{vec}(\tilde{z})$ , and then L2-norm is applied.

**Transformations.** Fig. 1 (right) shows the explored transformations in this work which fall into two groups, namely appearance and geometric transformations. In the context of gaze estimation, appearance and scale transformations do not change the 3D gaze direction label with respect to the camera coordinate system. In contrast, applying horizontal flip and rotation in image space results in respective changes in label space. Thus, for our proposed SwAT method, we only swap horizontal flip and rotation transformations. Further details of transformations can be found in the supplemental material (Sec. B.2).

### 3.3 Finetuning for Gaze Estimation

Gaze estimation is a regression task where the goal is to learn a mapping function  $\mathcal{H} : x \rightarrow g$  that maps the high-dimensional RGB images  $x \in \mathbb{R}^{H \times W \times 3}$  to low-dimensional 2D angles  $g \in \mathbb{R}^2$  i.e., yaw and pitch. The 2D angles are a compact representation of the 3D gaze direction vector in the camera coordinate system, the origin of which is the center of the face or the midpoint between the eyes, depending on the dataset.  $\mathcal{H}$  is a parameterized function, composed of a backbone encoder (e.g., ResNet) as well as a linear head (e.g., MLP). To perform gaze estimation, we first initialize the weights of the backbone with the pretrained weights previously learned through self-supervised pretraining. Then, the whole network is finetuned with gaze-annotated data using the L1 loss between the estimated angles  $\hat{g} = \mathcal{H}(x)$  and actual angles  $g$ , as follows (where  $N$  is the number of samples):

$$\mathcal{L}_{\text{gaze}} = \frac{1}{N} \sum_{i=1}^N \|g_i - \hat{g}_i\|_1, \quad (8)$$

## 4 Experiments and Results

In this section, we assess the performance of the proposed SwAT method through an exhaustive experimental evaluation to demonstrate the utility of the equivariance property under different scenarios. We refer the reader to supplemental material for robustness analysis (Sec. C), ablation studies and comparison with [E2] (Sec. D), and qualitative results (Sec. E).

## 4.1 Experimental Setting

**Datasets.** For the self-supervised pretraining stage, we use a curated dataset i.e., ETH-XGaze [43] without labels. It contains 756,540 images and 80 subjects for training, captured under controlled laboratory conditions. Since ETH-XGaze was specifically collected for the task of gaze estimation under controlled conditions, it is unclear whether the quality of unsupervised features remains the same while using an uncurated dataset. To shed light on this, we also use the VGG-Face dataset [30] for pretraining. VGG-Face is collected from the web, including 2,622 identities and about 1.5 M face images. For the finetuning phase, throughout various experiments, we use the publicly available Gaze360 [19] and MPIIFaceGaze [42] datasets, in addition to ETH-XGaze. Gaze360 is a physically unconstrained dataset collected in indoor and outdoor environments with a wide range of head poses. MPIIFaceGaze is a subset of the MPIIGaze [40] dataset, recorded while doing activities on the laptop.

**Implementation Details.** For the pretraining phase, we use SGD + LARS [38] optimizer with a batch size of 1024 distributed over 8 NVIDIA GeForce RTX 3090 GPUs. We pretrain for 100 epochs and experimentally found it to be sufficient. We use a weight decay of  $10^{-6}$  and the learning rate is set to 0.45 followed by an initial linear warmup stage for 10 epochs. Afterward, we use cosine learning rate decay [24] with a final value of 0.00045. As the encoder, we use ResNet [16] and a projection head that consists of a 2-layer MLP that maps the encoder output to 256-D. We experimentally set the number of prototypes  $M$  to 500. We perform the finetuning stage for 100 epochs using Adam optimizer [20], with a batch size of 512. We decay the learning rate at 40 and 80 epochs by 0.1. We set the input size to  $224 \times 224$  unless otherwise stated. Full details can be found in the supplemental material (Sec. A).

**Experimental protocol.** We use the dataset partitions provided by each dataset. A prior data normalization stage is commonly applied by creating a virtual camera with fixed intrinsic and extrinsic camera parameters, which reduces head pose variability and hence the training space [43]. However, this normalization may conceal the benefits of enforcing equivariance for geometric transformations, especially for already constrained datasets with little geometric variability. Furthermore, this stage cannot be applied accurately if camera parameters are not provided. Therefore, for the finetuning part of our methods (baselines and SwAT) we apply data normalization only to ETH-XGaze, since its test evaluation assumes normalized data, and to MPIIFaceGaze, to compare against previous approaches that performed the normalization stage. We also evaluate the unnormalized version of MPIIFaceGaze (referred to as MPIIFaceGaze\*) to better quantify the benefits of SwAT and compare its performance against the normalized counterpart. Throughout the paper, we use average angular gaze error in degrees to measure performance.

## 4.2 Evaluating the Unsupervised Features

After assessing the effectiveness of each individual transformation and finding an optimal composition for SwAV and SwAT (Sec. B.1, supplemental material), we evaluate the quality of unsupervised features. More precisely, the goal of this experiment is twofold: to explore whether the equivariance property provided by SwAT leads to a better representation compared to the invariance counterpart (SwAV), and to shed light on the quality of the unsupervised features with the curated (ETH-XGaze) and uncurated datasets (VGG-Face), used for pretraining. To do so, we perform a linear evaluation, where we freeze the backbone (ResNet-50) after pretraining and train a linear gaze regressor on top. Then, we measure the performance on the validation set that we manually create by splitting the available ETH-XGaze training set into training and validation sets. We also compare the unsupervised

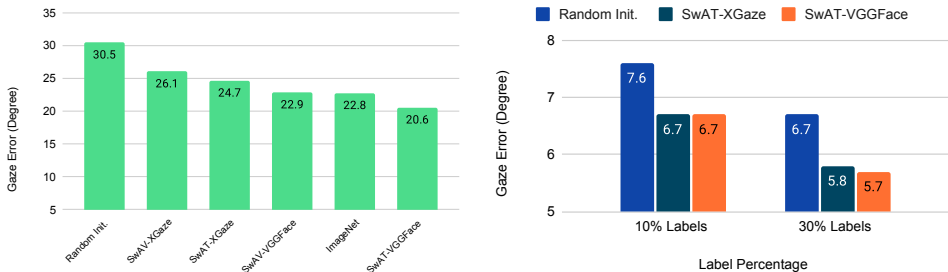


Figure 3: **Left.** Results of evaluating the unsupervised features of SwAV and SWAT pre-trained with ETH-XGaze and VGG-Face datasets compared to random and ImageNet-based initializations. Performance is measured on the manually created validation set of ETH-XGaze. **Right.** Results of semi-supervised learning using two subsets (10% and 30%) of the ETH-XGaze dataset at the subject level on the test set of ETH-XGaze.

features with ImageNet supervised features, which are widely used in current gaze estimation works as initialization.

Fig. 3 (left) shows the results of the linear evaluation on the validation set of ETH-XGaze. We can see that SWAT outperforms SwAV with both curated (ETH-XGaze) and uncurated (VGG-Face) datasets. More importantly, SWAT surpasses the supervised features pretrained on ImageNet, decreasing the gaze error from  $22.8^\circ$  to  $20.6^\circ$ . In the next experiments, we focus on comparing and evaluating SWAT in presence of labels for finetuning.

### 4.3 Semi-supervised Learning

In this evaluation, we examine the label-efficiency of SWAT. To achieve that, we perform semi-supervised learning on two subsets of the ETH-XGaze dataset. More precisely, we define two subsets i.e., 10% and 30% at subject level, and finetune the whole network on these subsets. As a baseline, we train a counterpart on the same subsets and with the same architecture but instead of using pretrained SWAT weights, we randomly initialize the weights. Fig. 3 (right) depicts the results of the semi-supervised learning. As can be seen, ResNet-50 pretrained with SWAT improves the baseline up to  $1.0^\circ$  when only 10% and 30% of labeled data at the subject level is available. This is of great importance in the gaze estimation context as recruiting fewer subjects saves cost and time.

### 4.4 Comparison to state of the art

We compare SWAT with state-of-the-art methods for full-face appearance-based gaze estimation. We pretrain SWAT with ResNet-50 as encoder on ETH-XGaze (without labels) and VGG-Face datasets. Then, we finetune the whole network using the aforementioned datasets. As a baseline, we also train the same encoder (ResNet-50) solely in a supervised fashion. Tab. 1 shows the comparison with the state of the art along with the datasets used for pretraining and the type of encoder. As can be seen, the supervised baseline is unable to outperform the state of the art, except on Gaze360. However, the same encoder boosted with SWAT unsupervised pretrained features achieves up to 25%, 5%, 14%, and 19% improvements compared to the supervised baseline on ETH-XGaze, Gaze360, MPIIFaceGaze, and MPIIFaceGaze\*, respectively. Furthermore, SWAT pretrained with the VGG-Face dataset outperforms SWAT pretrained on ETH-XGaze (without labels) on all four benchmarks. This suggests that SWAT can effectively make use of uncurated datasets. On ETH-XGaze, SWAT



Method	Pretrain	Arch.	ETH-XGaze	Gaze360	MPIIFace	MPIIFace*
Full-Face [14]	ImageNet	AlexNet+SW	N/A	N/A	<b>4.8</b>	N/A
Dilated-Net [8]	ImageNet	Dilated-CNN	N/A	N/A	<b>4.8</b>	N/A
RT-GENE [10]	ImageNet	VGG-16	N/A	N/A	<b>4.8</b>	N/A
Gaze360 [19]	ImageNet	ResNet-18	N/A	13.2	N/A	N/A
MTGLS [13]	MS-Celeb-1M	ResNet-50	N/A	12.8	N/A	N/A
ETH-XGaze [15]	ImageNet	ResNet-50	4.5	N/A	<b>4.8</b>	7.1 <sup>†</sup>
Wu et al. [16]	N/S	ResNet-18	N/A	13.2	N/A	N/A
Baseline (ours)	Random Init.	ResNet-50	5.9	12.2	5.7	8.5
SwAT (ours)	ETH-XGaze	ResNet-50	4.5	11.9	5.2	7.5
SwAT (ours)	VGG-Face	ResNet-50	<b>4.4</b>	<b>11.6</b>	5.0	<b>6.9</b>

Table 1: Comparison of SwAT with state-of-the-art full-face appearance-based gaze estimation works, reported as average angular gaze error (degrees). Best results are bolded. Performances of the state-of-the-art approaches are shown as reported by their authors, except values marked with <sup>†</sup>. MPIIFaceGaze\* denotes the unnormalized version of MPIIFaceGaze.

Method	Test		ETH-XGaze	Gaze360	MPIIFace	MPIIFace*
	Train					
Supervised	ETH-XGaze	-	-	30.0	23.5	17.5
	Gaze360	25.6	-	-	30.4	21.5
	MPIIFace	32.2	27.4	-	-	-
	MPIIFace*	35.5	28.9	-	-	-
SwAT	ETH-XGaze	-	-	<b>22.9</b>	<b>12.1</b>	<b>11.6</b>
	Gaze360	<b>19.4</b>	-	-	<b>13.0</b>	<b>12.8</b>
	MPIIFace	<b>29.5</b>	<b>24.9</b>	-	-	-
	MPIIFace*	<b>32.6</b>	<b>25.5</b>	-	-	-

Table 2: Comparison between supervised baseline and SwAT on cross-dataset evaluation. Numbers denote gaze error in degrees. Best results are bolded.

pretrained with VGG-Face outperforms the state of the art that utilizes the pretrained ImageNet supervised weights. In addition, SwAT improves the state of the art up to 9% on Gaze360 while slightly underperforming it on MPIIFaceGaze. However, we can better observe the benefit of SwAT on the unnormalized version of MPIIFaceGaze (MPIIFaceGaze\*), where SwAT improves the ETH-XGaze method with no data normalization by 0.2°. These results demonstrate the superior performance of SwAT in unrestricted scenarios.

## 4.5 Cross-dataset Evaluation

To evaluate the out-of-distribution generalization capability of SwAT, we perform a cross-dataset evaluation, i.e., training on a given dataset and testing on other datasets. We consider four datasets, namely, ETH-XGaze, Gaze360, MPIIFaceGaze, and MPIIFaceGaze\*. We use ResNet-50 as encoder, pretrained on VGG-Face using SwAT. We compare our self-supervised approach (SwAT) to a supervised baseline that is solely trained in a supervised fashion. Tab. 2 shows the results of cross-dataset evaluation. SwAT improves the supervised baseline by a large amount. In detail, SwAT achieves up to 24% relative improvement on the ETH-XGaze dataset, and outperforms the supervised counterpart by 24% on Gaze360, by 57% on MPIIFaceGaze, and by 41% on MPIIFaceGaze\*.

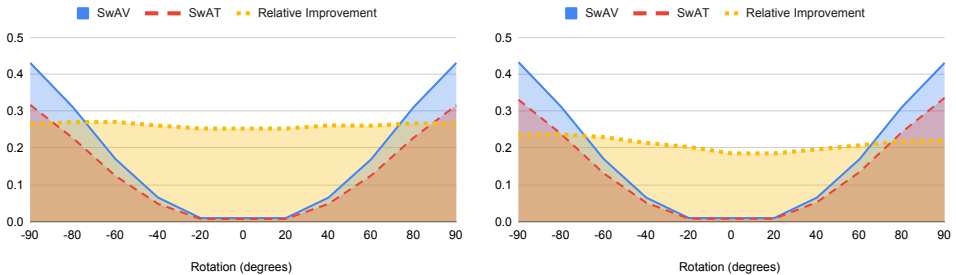


Figure 4: Results of calculating  $\mathcal{L}_{equ}$  for SwAV and SwAT on Gaze360 (Left) and MPIIFaceGaze\* (Right) datasets. The dotted lines shows the relative improvement achieved by SwAT over SwAV.

## 4.6 Equivariance Analysis

To evaluate the equivariance capability, we rely on the definition of equivariance (Eq. 3) and calculate the following metric ( $\mathcal{L}_{equ}$ ):

$$\mathcal{L}_{equ} = \frac{1}{N} \sum_{i=1}^N \|f_{\phi}(t_i^g(x_i)) - t_i^g(f_{\phi}(x_i))\|_2. \quad (9)$$

We compare  $f_{\phi}$  pretrained with SwAV and SwAT on the VGG-Face dataset. As the evaluation datasets, we specifically focus on unconstrained gaze scenarios and calculate  $\mathcal{L}_{equ}$  for Gaze360 and MPIIFaceGaze\*. We expect SwAT to achieve lower values, which indicates enforcing equivariance. Fig. 4 depicts the results of  $\mathcal{L}_{equ}$  on Gaze360 (left) and MPIIFaceGaze\* (right), varying rotation degrees. As shown, in both cases SwAT consistently outperforms SwAV in the whole rotation range. More precisely, on average, SwAT achieves 27% and 21% relative improvements compared to SwAV on Gaze360 and MPIIFaceGaze\*, respectively. Moreover, we calculate  $\mathcal{L}_{equ}$  for horizontal flip and find that SwAT improves SwAV by 26% on Gaze360 and 21% on MPIIFaceGaze\*.

## 5 Conclusion

In this paper, we explored the effectiveness of a self-supervised method in the context of gaze estimation, and proposed a novel approach (SwAT) to learn an equivariant representation for geometric transformations, i.e., rotations and horizontal flip. Our approach is task-agnostic and can be applied to any joint embedding-based self-supervised approach. We showed that SwAT learns more informative representations than other pretraining schemes for the task of gaze estimation. Our approach fueled by a large-scale uncurated dataset achieves more generalizable results, outperforming the supervised baselines and state-of-the-art approaches for both within- and cross-dataset settings. We also showed that our method achieves superior performance with fewer subjects. Thus, our approach can be leveraged to boost the performance of current gaze estimation systems in the real world via leveraging large-scale freely available face images on the Internet.

## Acknowledgements

This work has been partially supported by the Spanish project PID2019-105093GB-I00 and by ICREA under the ICREA Academia programme.

## References

- [1] Pulkit Agrawal, João Carreira, and Jitendra Malik. Learning to see by moving. In *ICCV*, 2015.
- [2] Adrien Bardes, Jean Ponce, and Yann LeCun. VICReg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR*, 2022.
- [3] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. In *NeurIPS*, 2020.
- [4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, 2020.
- [5] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *CVPR*, 2021.
- [6] Zhaokang Chen and Bertram E. Shi. Appearance-based gaze estimation using dilated-convolutions. In *ACCV*, 2018.
- [7] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *AAAI*, 2020.
- [8] Yihua Cheng, Xucong Zhang, Feng Lu, and Yoichi Sato. Gaze estimation by exploring two-eye asymmetry. In *IEEE Transactions on Image Processing*, 2020.
- [9] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *NeurIPS*, 2013.
- [10] Murthy L R D and Pradipta Biswas. Appearance-based gaze estimation using attention and difference mechanism. In *CVPR Workshops*, 2021.
- [11] Rumén Dangovski, Li Jing, Charlotte Loh, Seung-Jun Han, Akash Srivastava, Brian Cheung, Pulkit Agrawal, and Marin Soljaić. Equivariant contrastive learning. In *ICLR*, 2022.
- [12] Tobias Fischer, Hyung Jin Chang, and Y. Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *ECCV*, 2018.
- [13] Shreya Ghosh, Munawar Hayat, Abhinav Dhall, and Jarrod Knibbe. Mtgls: Multi-task gaze estimation with limited supervision. In *WACV*, 2022.
- [14] Spyros Gidaris, Praveer Singh, and Nikos Komodakis. Unsupervised representation learning by predicting image rotations. In *ICLR*, 2018.
- [15] Jean-Bastien Grill, Florian Strub, Florent Alché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, Bilal Piot, koray kavukcuoglu, Remi Munos, and Michal Valko. Bootstrap your own latent - a new approach to self-supervised learning. In *NeurIPS*, 2020.
- [16] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

- [17] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *CVPR*, 2020.
- [18] Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. A survey on contrastive self-supervised learning. In *ArXiv*, 2020.
- [19] Petr Kellnhofer, Adrià Recasens, Simon Stent, W. Matusik, and A. Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *ICCV*, 2019.
- [20] Diederick P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.
- [21] Rakshit Kothari, Shalini De Mello, Umar Iqbal, Wonmin Byeon, Seonwook Park, and Jan Kautz. Weakly-supervised physically unconstrained gaze estimation. In *CVPR*, 2021.
- [22] K. Krafska, A. Khosla, Petr Kellnhofer, Harini Kannan, S. Bhandarkar, W. Matusik, and A. Torralba. Eye tracking for everyone. In *CVPR*, 2016.
- [23] Xiao Liu, Fanjin Zhang, Zhenyu Hou, Li Mian, Zhaoyu Wang, Jing Zhang, and Jie Tang. Self-supervised learning: Generative or contrastive. In *IEEE Transactions on Knowledge and Data Engineering*, 2021.
- [24] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. In *ICLR*, 2017.
- [25] Ishan Misra and Laurens van der Maaten. Self-supervised learning of pretext-invariant representations. In *CVPR*, 2020.
- [26] Kenneth Alberto Funes Mora, Florent Monay, and Jean-Marc Odobez. Eyediap: a database for the development and evaluation of gaze estimation algorithms from rgb and rgb-d cameras. In *ETRA*, 2014.
- [27] Mehdi Noroozi and Paolo Favaro. Unsupervised learning of visual representations by solving jigsaw puzzles. In *ECCV*, 2016.
- [28] Cristina Palmero, Javier Selva, Mohammad Ali Bagheri, and Sergio Escalera. Recurrent cnn for 3d gaze estimation using appearance and shape cues. In *BMVC*, 2018.
- [29] Seonwook Park, Adrian Spurr, and Otmar Hilliges. Deep pictorial gaze estimation. In *ECCV*, 2018.
- [30] Omkar M. Parkhi, Andrea Vedaldi, and Andrew Zisserman. Deep face recognition. In *BMVC*, 2015.
- [31] Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.
- [32] Adrian Spurr, Aneesh Dahiya, Xi Wang, Xucong Zhang, and Otmar Hilliges. Self-supervised 3d hand pose estimation from monocular rgb via contrastive learning. In *ICCV*, 2021.
- [33] Yunjia Sun, Jiabei Zeng, Shiguang Shan, and Xilin Chen. Cross-encoder for unsupervised gaze representation learning. In *ICCV*, 2021.

- [34] Daniel E. Worrall, Stephan J. Garbin, Daniyar Turmukhambetov, and Gabriel J. Brostow. Interpretable transformations with encoder-decoder networks. In *ICCV*, 2017.
- [35] Yong Wu, Gongyang Li, Zhi Liu, Mengke Huang, and Yang Wang. Gaze estimation via modulation-based adaptive network with auxiliary self-learning. In *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.
- [36] Enze Xie, Jian Ding, Wenhai Wang, Xiaohang Zhan, Hang Xu, Peize Sun, Zhenguo Li, and Ping Luo. Detco: Unsupervised contrastive learning for object detection. In *ICCV*, 2021.
- [37] Yuyang Xie, Jianhong Wen, Kin Wai Lau, Yasar Abbas Ur Rehman, and Jiajun Shen. What should be equivariant in self-supervised learning. In *CVPR Workshops*, 2022.
- [38] Yang You, Igor Gitman, and Boris Ginsburg. Large batch training of convolutional networks. In *ArXiv*, 2017.
- [39] Yuechen Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. In *CVPR*, 2020.
- [40] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful image colorization. In *ECCV*, 2016.
- [41] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *CVPR*, 2015.
- [42] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It’s written all over your face: Full-face appearance-based gaze estimation. In *CVPR Workshops*, 2017.
- [43] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *ETRA*, 2018.
- [44] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. In *TPAMI*, volume 41, pages 162–175, 2019.
- [45] Xucong Zhang, Seonwook Park, Thabo Beeler, Derek Bradley, Siyu Tang, and Otmar Hilliges. Eth-xgaze: A large scale dataset for gaze estimation under extreme head pose and gaze variation. In *ECCV*, 2020.