# Towards Self-Supervised Gaze Estimation

A. Farkhondeh[1,2], C. Palmero[2,3], S. Scardapane[1], and S. Escalera[2,3]

[1]Sapienza University of Rome, [2]Computer Vision Center, [3]University of Barcelona

BMVC 2022

## Motivation

□ **Current supervised appearance-based gaze estimation methods cannot generalize well to novel distributions.** A possible solution: acquisition of larger in-the-wild, gaze-annotated datasets with more variability. However, collecting data with accurate gaze annotations is an unscalable and laborious process.

□ **An alternative solution: leveraging large-scale unlabeled face images using self-supervised learning (SSL).** However, current SSL methods [1] learn an invariant representation under appearance and geometric transformations. However, gaze estimation requires equivariance under geometric transformations.

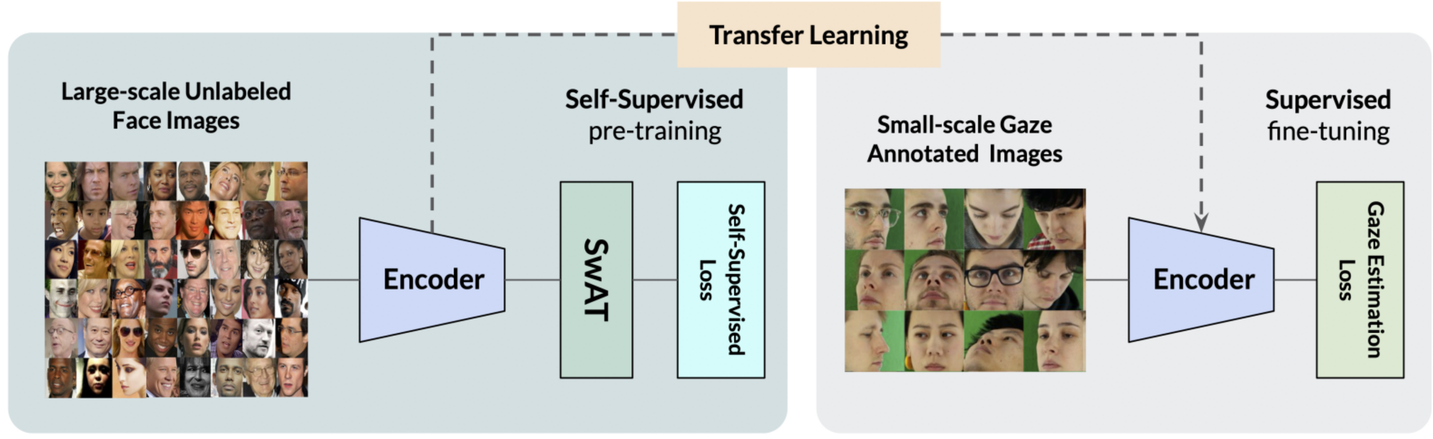



**Fig 1.** Overview. **Stage 1)** Self-Supervised Pre-training, **Stage 2)** Supervised Fine-tuning.

## Proposed Approach

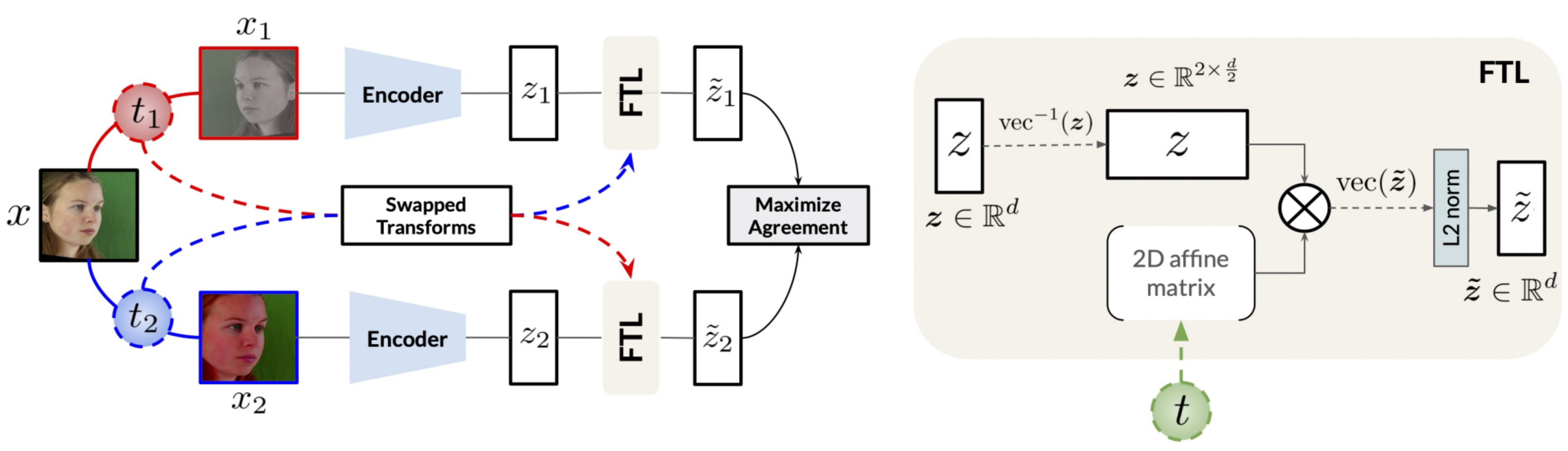



**Fig 2. Left.** SwAT overview. **Right.** Details of the feature transform layer (FTL).

### Self-Supervised Pretraining

□ **Pretext task:**
Maximize agreement between two differently transformed views of the same image.

□ **Maximizing agreement using SwAV [1]:** an online clustering-based method.
Swapped prediction of cluster assignments computed from vector representations.

$$\mathcal{L}_{\text{SwAV}} = \ell(\boldsymbol{z}_1, \boldsymbol{c}_2) + \ell(\boldsymbol{z}_2, \boldsymbol{c}_1)$$

### Equivariant Representation Learning

□ **SwAT: Swapping Affine Transformations**
1) Swap the affine transformations applied in image space, 2) Apply the swapped transformations to vector representations via feature transform layer, 3) Maximize agreement between transformation-equalized vectors.

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

□ **Feature Transform Layer (FTL)**
Feature-space equivalent of the image-space transformation.

### Fine-tuning for Gaze Estimation

1) Initialize CNN encoder with pre-trained weights of SwAT, 2) Attach a MLP head to regress gaze, 3) Fine-tune the whole network by minimizing L1 loss.

### Transformations

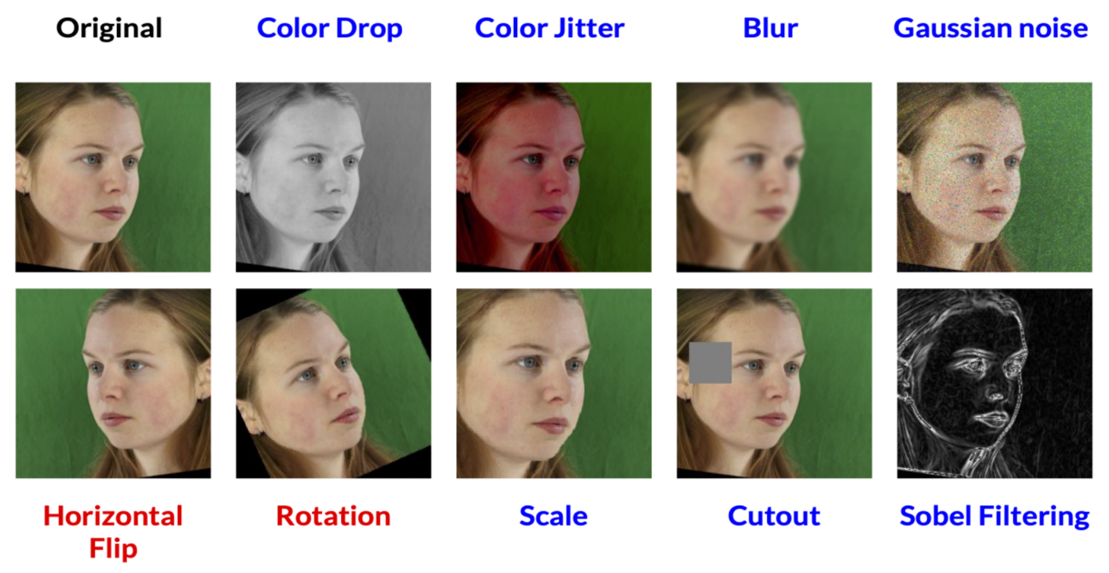



**Fig 3.** Explored transformations. Invariance vs. Equivariance.

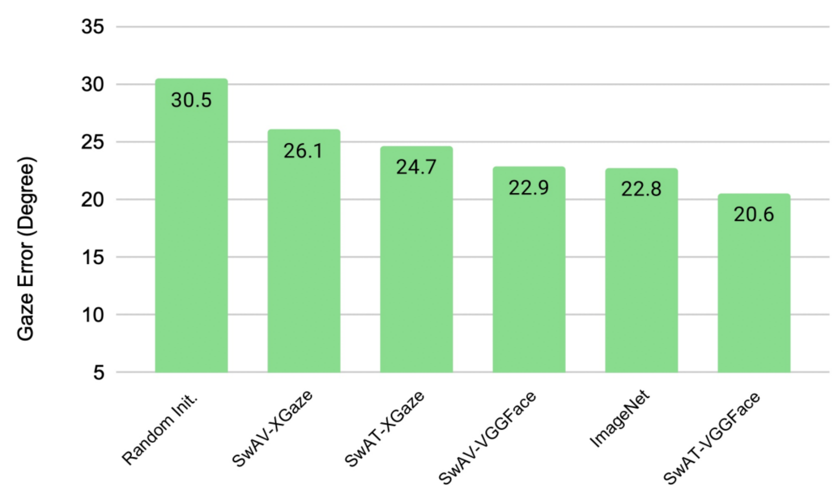
## Conclusions

□ SwAT Learns more informative representations than other pre-training schemes.
□ SwAT shows superior performance in low-data regimes.
□ SwAT outperforms the supervised baselines and state-of-the-art approaches for both within- and cross-dataset settings.

[1] 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.

## Experiments and Results

**Pre-training datasets**: ETH-XGaze (w/o lablels) and VGGFace
**Fine-tuning datasets:** ETH-XGaze, Gaze360, MPIIFace, and MPIIFace* (unnormalized)



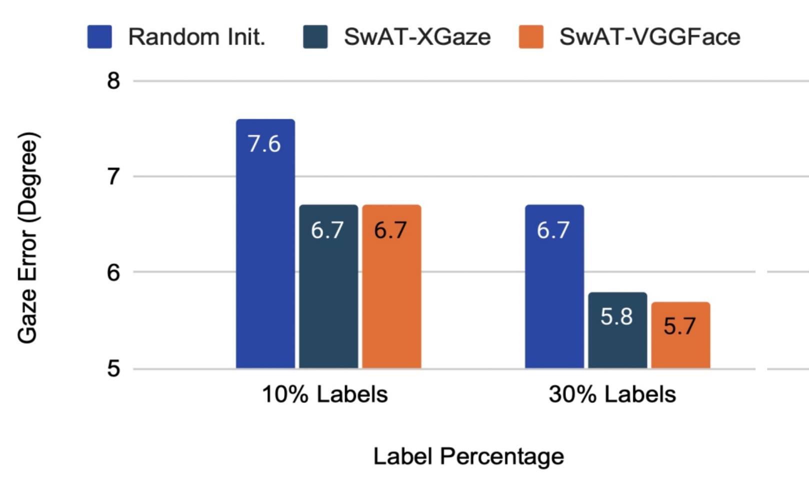




**Fig 4.** Results of evaluating the unsupervised features. **Fig 5.** Results of semi-supervised learning.

### Linear Probing (Fig.4)

□ SwAT-VGGFace achieves the lowest error compared to other pre-training schemes.
□ SwAT-VGGFace outperforms ImageNet supervised features.

### Semi-Supervised Learning (Fig.5)

□ SwAT achieves 1° less error compared to the supervised baseline when 10% and 30% of labels are used for fine-tuning.

### Comparison to State of the Art

| Method | Pretrain | Arch. | ETH-XGaze | Gaze360 | MPIIFace | MPIIFace* |
|---|---|---|---|---|---|---|
| Full-Face [42] | ImageNet | AlexNet+SW | N/A | N/A | **4.8** | N/A |
| Dilated-Net [6] | ImageNet | Dilated-CNN | N/A | N/A | **4.8** | N/A |
| RT-GENE [12] | ImageNet | VGG-16 | N/A | N/A | **4.8** | 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 [45] | ImageNet | ResNet-50 | 4.5 | N/A | **4.8** | 7.1† |
| Wu et al. [35] | 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 | **4.4** | **11.6** | 5.0 | **6.9** |

**Tab 1.** Comparison to state of the art.

□ SwAT outperforms the supervised baseline on all benchmarks (up to 25%).

□ SwAT achieves SoTA results on ETH-XGaze, Gaze360, and MPIIFace*.

### Cross-dataset Evaluation

| Method | Train \ Test | ETH-XGaze | Gaze360 | MPIIFace | MPIIFace* |
|---|---|---|---|---|---|
| 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 | - | **22.9** | **12.1** | **11.6** |
| | Gaze360 | **19.4** | - | **13.0** | **12.8** |
| | MPIIFace | **29.5** | **24.9** | - | - |
| | MPIIFace* | **32.6** | **25.5** | - | - |

**Tab 2.** Cross-dataset evaluation.

□ SwAT outperforms the supervised baseline on all benchmarks.

□ SwAT achieves up to 57% relative improvement.

### Equivariance Analysis


**Fig. 6.** Equivariance analysis on Gaze360.

□ For rotation, on average, SwAT achieves 27% relative improvement compared to SwAV.

□ For horizontal flip, SwAT improves SwAV by 26%.

$$\mathcal{L}_{equ} = \frac{1}{N}\sum_{i=1}^{N} ||f_\phi(t_I^g(\boldsymbol{x}_i)) - t_F^g(f_\phi(\boldsymbol{x}_i))||_2$$