Self-Supervised Pretraining

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.

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.

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

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.

CONDITIONED TRANSFORMATIONS
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.

¨ Pretext task: Maximize agreement between two differently transformed views of the same image.
¨ Maximizing agreement using SwAV [1]: an online clustering-based method.

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

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.

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.

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

Pretext task: Maximize agreement between two differently transformed views of the same image.
Maximizing agreement using SwAV [1]: an online clustering-based method.

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.

Towards Self-Supervised Gaze Estimation
A. Farkhondeh1,2, C. Palmero2,3, S. Scardapane1, and S. Escalera2,3
1Sapienza University of Rome, 2Computer Vision Center, 3University of Barcelona

Proposed Approach

Self-Supervised Pretraining

Equivariant Representation Learning

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

Experiments and Results

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

Linear Probing (Fig.4)
□ SwAT-VGGFace achieves the lowest error compared to other pre-training schemes.
□ SwAT-VGGFace outperforms ImageNet supervised features.

Comparison to state of the art

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.

Cross-dataset Evaluation

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.

Cross-dataset Evaluation

SwAT outperforms the supervised baseline on all benchmarks (up to 25%).
SwAT achieves SoTA results on ETH-XGaze, Gaze360, and MPIIFace*.

SwAT outperforms the supervised baseline on all benchmarks.
SwAT achieves up to 57% relative improvement.

Equivariance Analysis

For rotation, on average, SwAT achieves 27% relative improvement compared to SwAV.
For horizontal flip, SwAT improves SwAV by 26%.