Enhancing 3D Gaze Estimation in the Wild using Weak Supervision with Gaze Following Labels

Pierre Vuillecard   Jean-Marc Odobez

Idiap Research Institute, Switzerland

Ecole Polytechnique Fédérale de Lausanne, Switzerland
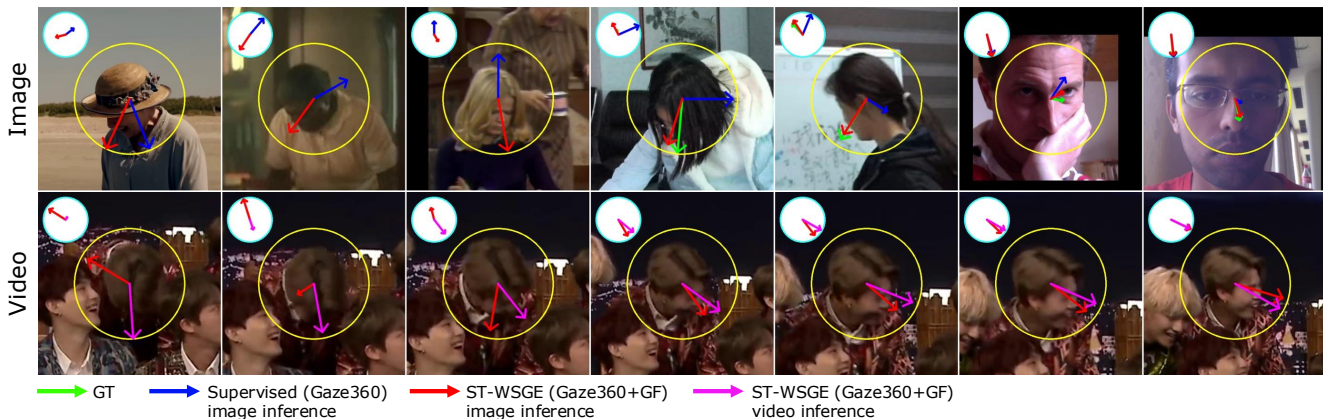
{ pvuillecard, odobez}@idiap.ch

Figure 1. **Significance of ST-WSGE.** Our self-training based weakly-supervised framework for robust 3D gaze estimation in real-world conditions (e.g., varying appearance, extreme poses, resolution, and occlusion). All predictions used our image and video agnostic Gaze Transformer (GaT) model. Top row: importance of the training diversity using ST-WSGE and GazeFollow (GF) for generalization compared to standard supervised methods. Bottom row: influence of temporal context between image and video inference. Circles in images represent unit disks where 3D gaze vectors are projected onto the image plane (x, y in yellow) and a top-down view (x, z in blue). Images from VideoAttentionTarget, GFIE, and MPIIFaceGaze datasets.

Abstract

Accurate 3D gaze estimation in unconstrained real-world environments remains a significant challenge due to variations in appearance, head pose, occlusion, and the limited availability of in-the-wild 3D gaze datasets. To address these challenges, we introduce a novel Self-Training Weakly-Supervised Gaze Estimation framework (ST-WSGE). This two-stage learning framework leverages diverse 2D gaze datasets, such as gaze-following data, which offer rich variations in appearances, natural scenes, and gaze distributions, and proposes an approach to generate 3D pseudo-labels and enhance model generalization. Furthermore, traditional modality-specific models, designed separately for images or videos, limit the effective use of available training data. To overcome this, we propose the Gaze Transformer (GaT), a modality-agnostic architecture capable of simultaneously learning static and dynamic gaze information from both image and video datasets. By combining 3D video datasets with 2D gaze target labels from gaze following tasks, our approach achieves the fol-

lowing key contributions: (i) Significant state-of-the-art improvements in within-domain and cross-domain generalization on unconstrained benchmarks like Gaze360 and GFIE, with notable cross-modal gains in video gaze estimation; (ii) Superior cross-domain performance on datasets such as MPIIFaceGaze and Gaze360 compared to frontal face methods. Code and pre-trained models will be released to the community.

1. Introduction

Non-verbal behaviors play a crucial role in human communication, often conveying more information than words alone. Among the various forms of non-verbal cues, eye gaze stands out as an important signal for understanding human behavior, including attention, communication, intents, and mental state. Consequently, gaze signals have been used in many applications. Some applications require accurate gaze for frontal head pose such as AR/VR [5], 3D avatar animation [45], human-computer interaction [3, 44], and driver behavior monitoring [27]. While other applications focus on robust 3D gaze estimation from a wide range

of head poses such as medical and psychological analysis [29] or human-robot interaction [2, 30, 48].

In this paper, our goal is to develop a robust 3D gaze estimation for in-the-wild applications with unconstrained head pose and real-world environments. In the literature, this refers to the less explored and highly challenging problem of "physically unconstrained gaze estimation" [28].

Motivations. Estimating gaze in unconstrained, real-world settings poses unique challenges not fully addressed by current lab-based datasets, which are primarily collected in controlled screen-target setups [18, 19, 32, 42, 60, 63]. While these datasets have enabled recent approaches to achieve high accuracy in frontal 3D gaze estimation from monocular images [1, 7, 9–11, 18, 55, 61], their effectiveness declines in real-world scenarios. This limitation stems from restricted gaze distributions, lab-specific conditions, limited subject diversity, and reliance on potentially noisy head-pose estimates for normalizing eye and face images [62].

To address the lack of data for unconstrained gaze estimation, Gaze360 [28] and GFIE [25] were developed. Although these datasets have advanced the field, models trained on them continue to struggle with challenging, real-world conditions (see Fig. 1), particularly when facing extreme head poses, partial eye occlusions, varying resolutions from diverse camera-to-subject distances, and a wide range of appearances (e.g., skin tones, hairstyles, facial expressions). This limitation is largely due to insufficient diversity in the training data, as collecting high-quality, naturally occurring, and diverse 3D gaze data is complex, resource-intensive, and not easily scalable.

To overcome these limitations, researchers have explored using "secondary" labels that are easier to obtain, *e.g.* by relying on internet data. For example, Kothari *et al*. [31] leveraged 2D gaze direction labels from the "Looking at Each Other" (LAEO) dataset [38]. By applying geometric constraints and head-size heuristics, they generated pseudo-3D gaze data. While this approach showed some generalization improvement, the authors noted that LAEO's gaze distribution is primarily horizontal and requires images containing at least two people with mutual gaze, which limits sample diversity and availability.

Here we aim to utilize more general 2D gaze annotations from the gaze following task [14, 43, 50, 51]. Although the ground truth for gaze following is defined as the 2D pixel location a person in the scene is looking at, we can repurpose it as 2D gaze direction ground truth. Compared to LAEO, gaze following datasets offer greater diversity, with broader gaze distributions and a wider variety of natural scenes.

In addition, in contrast to [31], we propose a Self-Training Weakly-Supervised Gaze Estimation (ST-WSGE) framework using a two-stage training approach without relying on heuristics or relative depth estimation to generate pseudo-3D gaze labels. First, we train a gaze network on existing 3D gaze datasets. We then use this network's predictions on gaze-following data, combined with 2D gaze ground truth, to create 3D gaze pseudo-labels. In the second stage, we retrain a similar gaze network using both 3D gaze datasets and gaze-following datasets with these pseudo-labels. Our approach minimizes the need for labor-intensive, unconstrained 3D gaze labeling and demonstrates significant improvements over state-of-the-art methods in both within-domain and cross-domain generalization on Gaze360, GFIE, and MPIIFaceGaze [60].

Given the scarcity of in-the-wild 3D gaze datasets, another challenge lies in how to leverage both image and video data effectively. Modality-specific models restrict the training set to modality-specific datasets, limiting their ability to benefit from all available resources. While static models can draw on large image datasets such as GazeFollow [14], temporal dynamics are also essential for robust 3D gaze estimation in unconstrained environments [28, 40], and is particularly valuable when the eye region is obscured, whether due to occlusions, low resolution, or blinking (see Fig. 1).

To address this, one approach is to pre-train a model on images and transfer the weights to a temporal model using techniques like filter inflation, where 2D filters are extended to 3D models, as done in prior adaptations for video tasks [6]. However, this method is more suited to fine-tuning and risks catastrophic forgetting, where the model loses pre-trained knowledge [39]. Alternatively, images can be duplicated to simulate fixed-length video clips, allowing for training on both image and video datasets in a temporal model. However, this can impair the learned gaze dynamics, as synthetic clips lack genuine motion information.

Transformers offer a promising solution for handling multiple modalities. Inspired by recent work [20, 21], we propose a Gaze Transformer (GaT) designed to encode both image and video inputs into a shared representation. This allows us to leverage labeled datasets more effectively, by training jointly on image and video data. We demonstrate better cross-modal generalization, and that using image datasets enhances video gaze prediction, thus enabling a more versatile and robust 3D gaze estimation model.

Contributions. They can be summarized as:
- **ST-WSGE, a novel learning framework enhancing generalization.** To address the lack of diverse, naturalistic 3D gaze datasets, we leverage 2D gaze-following datasets using 3D pseudo labels. Combining these with 3D gaze datasets in a two-stage manner, we demonstrate improved 3D generalization on several benchmarks.
- **A visual modality agnostic Gaze Transformer (GaT) architecture making efficient use of existing gaze datasets.** By allowing simultaneous learning from 3D gaze image and video datasets, it outperforms modality-specific models, resulting in better static and dynamic

gaze representations, better capturing spatiotemporal patterns in head sequences compared to the state-of-the-art.

- **State-of-the-art results.** Our approach surpasses existing methods in both unconstrained (Gaze360, GFIE) and constrained (MPIIFaceGaze) environments, achieving superior results in within- and cross-dataset evaluations.

These contributions position our approach as ideal for real-world unconstrained 3D gaze estimation applications.

2. Related Work

Our research pertains to 3 main aspects: unconstrained gaze estimation, temporal gaze modeling, and generalization using additional data and labels to bridge the domain gap between controlled setups and real-world data.

Unconstrained Gaze Estimation. Most 3D gaze estimation models address the frontal face gaze prediction task [9, 10, 18, 33, 49, 58, 61], relying on normalized frontal face crop as input. These methods tend to fail under partial occlusion of the eyes due to extreme head pose. Nevertheless, at 90-135 head pose yaw, a significant part of one eyeball is still often visible and informative for gaze estimation [28]. For this reason, few works tackle the most challenging setting of "physically unconstrained gaze estimation" without constraint on the head pose. Kellnhofer *et al*. [28] are the first to collect a physically unconstrained 3D gaze dataset Gaze360 and develop a method that used head crop as input. Then, combining different head crop scales proved to be beneficial [8] since more resolution helps on the frontal face while more context is beneficial for profiles and back heads. Following this idea, MCGaze [22] used a spatiotemporal interaction module between head, face, and eye features in an end-to-end manner to extract local eyes and global head features. These approaches focus on within-data performance, while in this work we aim to improve both within and the generalization as discussed in the following section.

Generalization in the Wild. Bridging the dataset's domain gap challenge is crucial for 3D gaze estimation in real-world applications. Two trends have been explored to adapt to specific target domains effectively: One leverages few labeled samples, while the other uses only unlabeled samples [13, 26, 28, 53, 54, 57]. In contrast, gaze generalization models focus on enhancing cross-domain performance without any prior knowledge of the target domain. For instance, the methods proposed in [4, 13, 54] demonstrate improved generalization by learning robust general features (*e.g.* via image rotation consistency) for gaze estimation across varying conditions. Even if those methods focus on constrained settings with face crop as input, we compare our approach with them to show the effectiveness in frontal face generalization.

Furthermore, to improve in-the-wild generalization, researchers seek to exploit diverse weak gaze labels that can

be easily or automatically generated on in-the-wild data. In this direction, Zhang *et al*. [59] automatically generates a new 3D gaze dataset, MPSGaze, by blending on images of people from the Widerface datasets [56], eyes from images of the ETH-Xgaze dataset with known 3D gaze and similar head pose. While this greatly improves diversity with more than 10000 new identities, this method generates only near frontal faces and might impact the appearance of the face. In another study, Ververas *et al*. [53] used eyeball fitting techniques to create pseudo-3D gaze on new face datasets. They improved generalization, but their work is also restricted to frontal faces. Finally, Kothari *et al*. [31] used a weakly-supervised learning framework for improved generalization using pseudo 3D gaze labels from 2D gaze LAEO labeled datasets. However, as acknowledged by the authors, the 2D gaze distribution of LAEO is limited horizontally. In our work, we follow this idea but leverage a more diverse gaze distribution and natural scene 2D gaze label obtained from the annotation of where people look in the scene. Using a self-training learning approach with generated 3D pseudo labels via geometric projection, we show improved within and cross-dataset generalization on unconstrained Gaze360 and GFIE [25] and frontal MPIIFaceGaze [60] datasets.

Dynamic 3D Gaze Estimation has not been extensively explored due to the lack of available datasets. EYE-DIAP [19] and EVE [42] are video datasets collected in constrained settings, resulting in mostly frontal poses. In this particular context, Park *et al*. [42] and Palmero *et al*. [41] estimated the gaze from face crop image sequences but only showed marginal improvement compared to static methods. Indeed, it is questionable if eyeball dynamics have temporal dependencies besides the ones due to specific tasks or scenarios (*e.g.* reading). Nevertheless, in unconstrained settings with low resolution and head pose dynamic scenarios, temporal methods show benefits in encoding the head and eye dynamics [8, 22, 23, 28, 40]. For instance, seen from a far distance, head and body orientation dynamic revealed to be an important prior for gaze estimation when eyes are barely visible [40]. Unconstrained video gaze data is challenging to collect. Beyond Gaze360, GFIE [25] is the only other 3D gaze dataset for gaze following, using complex laser setups, yet it is limited to indoor settings and lacks natural scene and gaze dynamics. The scarcity of video 3D gaze datasets hampers the development of video-based methods that generalize to real-world data. To address this, we propose a unified model trained on both image and video datasets, demonstrating improved video prediction through diverse, large-scale data.

Furthermore, current video-based gaze estimation methods typically employ a backbone to extract features from image sequences, followed by a Recurrent Neural Network
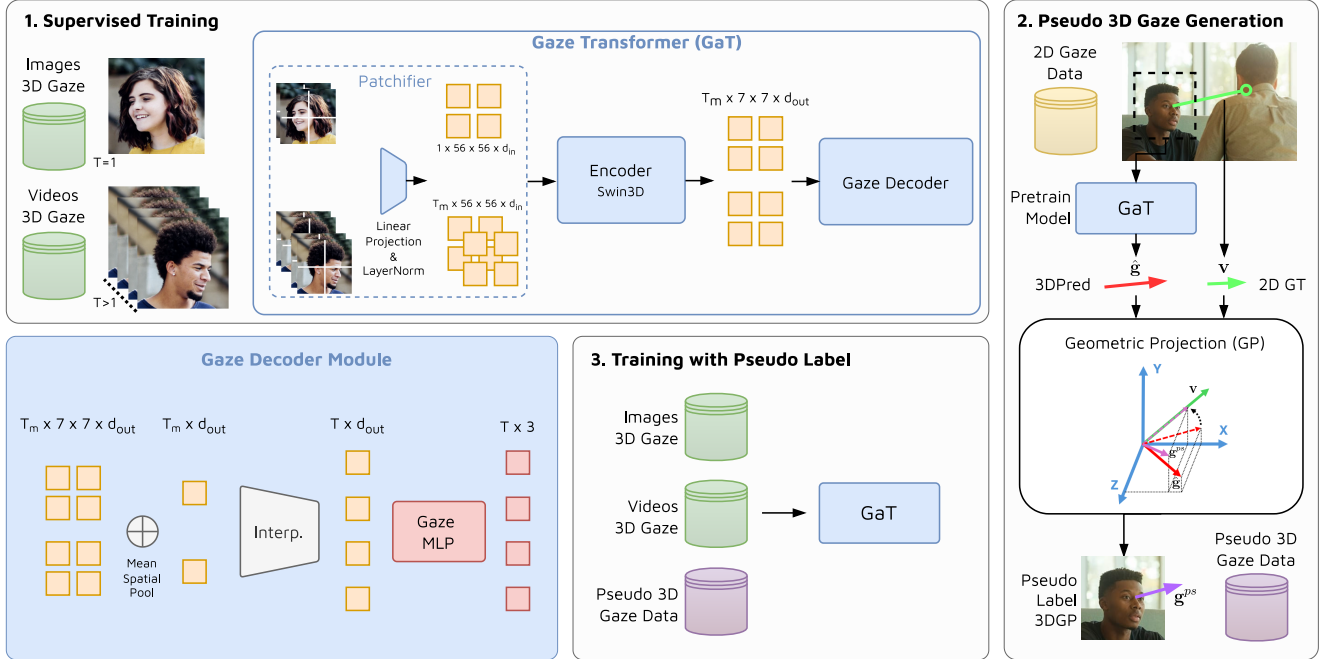
Figure 2. **Our ST-WSGE training framework.** 1. In the first stage, we train a Gaze Transformer (GaT) on both image and video 3D gaze datasets. 2. Using the trained network, 3D gaze is inferred on 2D gaze dataset. Then, a geometric rotation is applied to generate a pseudo 3D gaze label from the inferred 3D gaze that is aligned to the 2D ground truth gaze label in the image plane. 3. In the second stage, we train a similar gaze network as in 1. using available 3D gaze datasets and gaze following datasets with 3D pseudo labels.

(RNN) to capture temporal dynamics [8, 26, 28, 31, 41, 42]. However, these approaches do not explicitly model the spatiotemporal interactions in the input sequence. To address this limitation, we investigate a spatio-temporal model to encode subtle eye motion or head pose changes in the input sequence directly.

3. ST-WSGE Method

3.1. Self-Training Pipeline

We propose a two-stage training pipeline for gaze estimation to leverage any 2D gaze datasets, as presented in Fig. 2. In the first stage, a gaze network is trained on image and video 3D gaze datasets in a supervised manner. Next, the network is used to infer 3D gaze on 2D gaze datasets. Since only the gaze's depth is missing in the 2D gaze label, we employ a geometric transformation to generate a robust pseudo-3D gaze label from the inferred 3D gaze that is aligned with the 2D gaze label. We assume that a model pre-trained on unconstrained 3D gaze datasets provides a good prior z-estimation. In the second stage, a similar gaze network is trained in a supervised regime using both gaze following data with 3D pseudo labels and 3D gaze datasets.

3.2. Gaze Transformer (GaT)

Model Architecture Motivation. Accurate and robust 3D gaze estimation in the wild requires three key capabilities:

capturing fine local details from the eye region when visible; extracting global information from head orientation which is particularly useful when the eyes are occluded or partially obfuscated; capturing small motion of head pose and eyes in the temporal domain to capture subtle gaze shifts. A model capable of training on both image and video data is especially valuable, as it broadens the range of available training datasets. Convolutional Neural Networks (CNNs) excel in frontal gaze estimation due to their ability to extract local eye features [1, 28, 55] but may encounter more difficulty in global reasoning (i.e. merging pose and eye information), and extending CNNs to handle temporal data within a single modality-agnostic model is challenging. Vision Transformers (ViTs) [16], as noted by Cheng *et al.* [10], are less suited for gaze estimation since they may miss critical local details, especially when the eye region is split across multiple patches. In contrast, hierarchical transformer architecture [35, 46] offers a flexible architecture to capture both local and global features. For instance, the Swin Transformer, which uses smaller patches (typically 4×4 vs 16×16 in standard ViTs), is better able to capture fine local details. Its "shifted window" mechanism, which applies self-attention within local windows that shift at regular intervals, effectively aggregates local and global context. Extending the Swin Transformer to the temporal dimension has proven successful for temporal tasks on sev-

eral benchmarks [36]. Additionally, transformers are versatile, recent work has demonstrated their effectiveness when trained on both image and video datasets within a single model [20, 21]. Inspired by these approaches, we introduce our Gaze Transformer, GaT, with several modifications for 3D gaze estimation, as illustrated in Fig. 2 and detailed in the following sections.

Patchifier. The model needs a common representation format to encode both image and video input. Following [16, 17, 20, 35, 52], images and videos are represented as 4D tensors $\mathbf{X} \in \mathbb{R}^{T \times H \times W \times 3}$, with $T = 1$ for an image $\mathbf{I}$, and $T > 1$ for a video clip $\mathbf{V}$. Then, the input $\mathbf{X}$ is divided into a collection $\{\mathbf{x}_i\}^N$ of 4D sub-tensor (patches) $\mathbf{x}_i \in \mathbb{R}^{t \times h \times w \times 3}$, as presented in Fig. 2. Following [20, 21, 38, 52], we use $t = 2$. When working with image only, we duplicate the image instead of zero-padding because we find better cross modalities generalization from video to image. Then, a shared linear layer and LayerNorm are applied to project the patches to a token representation.

Encoder. The tokens from the patchifier are then fed into a tiny Swin3D hierarchical spatiotemporal encoder. It relies on self-attention within nearby tokens in a spatiotemporal window that is shifted every time. In addition, it uses two sets of relative positional encoding: one spatial and one temporal. Because of the hierarchical representation, the number of tokens is reduced by patch merging layers as the network gets deeper. The temporal output dimension is reduced by a factor of two. The output tokens are then fed to a gaze decoder module.

Gaze Decoder. We first apply a mean spatial pooling on the output tokens, followed by an interpolation function to double the temporal dimension to match the input length (for images, interpolation is skipped). Finally, a shared MLP with a single hidden layer is applied to each token to predict a normalized 3D gaze vector.

Baseline Networks. Different approaches exist to process image and video in a single model. To compare the performance of our GaT model, we develop in addition two baselines. Given that the static Swin(2D) transformer gives good performance on gaze estimation. We add a temporal encoder to model the gaze dynamic. Therefore, we develop Swin(2D)-LSTM which first processes a set of images using Swin and outputs a set of embedding for each image. Then, it is fed to a bidirectional LSTM followed by a shared gaze MLP on each output to produce a gaze vector. Similarly, the second baseline called Swin(2D)-Tr replaces the LSTM by a transformer. The output tokens from each image are projected to a lower dimension followed by a LayerNorm and absolute spatiotemporal encoding. Then, the transformer is applied to the spatiotemporal output token. Finally, a spatial mean pooling is applied followed by a similar gaze MLP. Both architectures are input agnostic and are compared in an ablation study.

3.3. Pseudo 3D Gaze Generation

Creating 2D gaze datasets is easier than creating 3D gaze datasets, as annotation can be completed after the images have been collected, unlike 3D gaze which can not be annotated by humans and require a special setup. As a result, 2D datasets like GazeFollow offer a broad gaze, head pose distribution, and head/face appearance diversity. There are different possibilities to leverage such a dataset which will be discussed in the ablation sections. For instance, Kothari *et al.* [31] used 2D gaze from LAEO labels and 3D fitted head models for z-direction estimation. In contrast, as presented in Fig. 2, our 3D pseudo gaze generation method assumes that a pre-trained model trained on unconstrained 3D gaze datasets can provide a good prior z-estimation, which is confirmed by our experiments. Combining the z component of predicted 3D gaze with 2D gaze ground truth provides a robust pseudo 3D gaze label. Then, using this pseudo label as an additional label during training, we report improvement in unconstrained generalization.

Geometric Projection. The predicted 3D Gaze (3DPred) $\hat{\mathbf{g}}$ and the 2D ground truth (2D GT) $\mathbf{v} = (v_x, v_y)$ are combined such that the image plane projection of the pseudo 3D gaze (3DGP) $\mathbf{g}^{ps}$ is aligned with the 2D ground truth $\mathbf{v}$. Therefore, a rotation is applied to $\hat{\mathbf{g}}$ around the z-axis such that $\mathbf{g}^{ps}$ has the same x,y direction as $\mathbf{v}$:

$$\mathbf{g}^{ps} = (v_x \|(\hat{g}_x, \hat{g}_y)\|_2, v_y \|(\hat{g}_x, \hat{g}_y)\|_2, \hat{g}_z) \qquad (1)$$

3.4. Training Strategy

In both training stages, the objective is to train our GaT model on a collection of both image and video datasets with gaze label $\{(\mathbf{X}, \mathbf{g})_j\}$ where $\mathbf{g} \in \mathbb{R}^{T \times 3}$ with T=1 for images, which creates different training challenges to be addressed.

Video Training Data. As our model is modality agnostic, a video dataset can be considered both as a set of video clips or as a collection of images. These views of the data are not equivalent, as, typically, considering the data as video-clip at training will impact more inference on videos at test time rather than on images. Hence, a video dataset can be used twice as an image or video training dataset[1]. We will see in ablations that it can impact the modality generalization capability.

Mini-batch Strategy. Different mini-batch strategies have been proposed in the literature to handle multiple datasets. One approach mixes samples from each of the datasets, but in our case, this requires careful implementation because images and videos don't have the same dimensions. Another strategy creates batches from one dataset at a time and alternates between them. This approach has proven effective in previous work [20, 24] and we followed

---

[1]By convention, when reporting experiments, for video datasets, we will add the suffix I when it is considered as an image dataset, V in the video case, and I&V when it is used twice as image and as video dataset.

Method	Training Data	GF Additional Label				G360 Full		GFIE		MPII	EDIAP
		None	2D	3DPred	3DGP	Img	Vid	Img	Vid	Img	Vid
Supervised	G360I&V	✓				13.6	12.6	21.9	20.9	7.4	8.3
Weakly-Sup (WS)	G360I&V+GF		✓			**13.1**	**12.1**	16.1	15.7	6.5	9.2
Self-Training (ST)	G360I&V+GF			✓		13.6	12.7	20.2	19.7	7.4	**7.7**
ST-WSGE	G360I&V+GF				✓	13.2	12.2	15.9	15.5	6.4	8.2
Supervised	GFIEI&V	✓				30.6	29.9	15.7	15.4	23.8	37.8
Weakly-Sup (WS)	GFIEI&V+GF		✓			22.9	22.1	12.5	12.2	24.4	33.0
Self-Training (ST)	GFIEI&V+GF			✓		29.7	29.4	14.9	14.9	21.2	34.6
ST-WSGE	GFIEI&V+GF				✓	**21.5**	**21.1**	13.0	12.7	17.3	16.7

Table 1. **Ablation study for the self-training weakly-supervised learning framework.** We experiment with our GaT model, two 3D gaze datasets Gaze360 and GFIE, and three ways to include GazeFollowing labels (GF). The **best** and the second best scores are in bold and underlined, respectively.

Model	Training modality		G360 Full		G360 180		G360 40	
	Img	Vid	Img	Vid	Img	Vid	Img	Vid
Swin(2D)-LSTM	✓		14.33	13.97	12.17	11.89	9.73	9.47
Swin(2D)-LSTM		✓	14.75	13.05	12.58	10.98	9.86	8.67
Swin(2D)-LSTM	✓	✓	13.93	13.02	11.76	10.94	8.88	8.27
Swin(2D)-Tr	✓		14.05	14.53	12.02	12.78	9.63	9.32
Swin(2D)-Tr		✓	14.05	12.63	12.05	10.54	9.46	8.14
Swin(2D)-Tr	✓	✓	13.81	12.91	11.96	11.09	9.43	8.67
GaT	✓		13.95	13.82	11.95	11.78	9.58	8.89
GaT		✓	13.87	12.31	11.89	10.39	9.29	7.95
GaT	✓	✓	13.64	12.60	11.66	10.67	9.10	8.23

Table 2. **Ablation study for the gaze model network.** Since different models are image and video training agnostic, we experiment with three models on three training modalities dataset combinations using Gaze360 as the training set.

Training Data	G360 Full		G360 180		G360 40		GFIE	
	Img	Vid	Img	Vid	Img	Vid	Img	Vid
G360V+GF	13.5	12.1	11.6	10.2	8.3	7.7	15.7	17.9
G360I&V+GF	13.2	12.2	11.3	10.3	8.6	7.7	15.9	15.5
GFIEV+GF	22.8	24.2	22.4	23.9	29.9	31.8	13.4	13.0
GFIEI&V+GF	21.5	21.1	20.6	20.3	26.6	26.7	13.0	12.7

Table 3. **Impact of the training datasets modalities on cross-modal generalization.** We experiment with GaT model, ST-WSGE framework, and different training dataset modalities.

this approach here. In addition, dataset size imbalance is another challenge, as dataset sizes range from 30k to 120k samples. To address this, we balance the datasets by oversampling smaller ones and undersampling larger ones so that each dataset contributes equally during an epoch.

Loss. For training the model, we utilize a temporal weighted average of the angular loss, which represents the angular difference between the predicted gaze vector $\hat{\mathbf{g}}$ and the ground truth $\mathbf{g}$ in degree:

$$\mathcal{L}_{gaze}(\hat{\mathbf{g}}, \mathbf{g}) = \frac{1}{T} \sum_{t=1}^{T} \frac{180}{\pi} \arccos(\frac{\hat{\mathbf{g}}_t^T \mathbf{g}_t}{\|\hat{\mathbf{g}}_t\| \|\mathbf{g}_t\|}) \quad (2)$$

4. Experiments

4.1. Datasets

In this work, we employ five 3D gaze datasets: two video unconstrained datasets for training and evaluation: Gaze360 (G360) [28], GFIE [25], and three constrained only for generalization MPSGaze (MPS) [59], MPIIFaceGaze (MPII) [60] and EYEDIAP (EDIAP) [19] (EDIAP), with only EDIAP being a video dataset. As shown in Fig. 3,
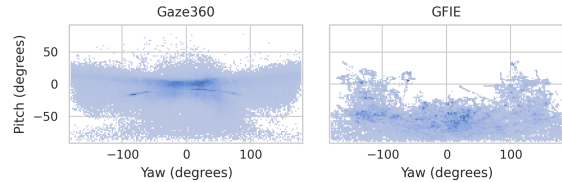

Figure 3. **Dataset gaze distribution.** Gaze in polar coordinates.

G360 and GFIE differ considerably in their gaze distribution, which makes cross-dataset evaluations challenging. In addition, we consider the 2D gaze following dataset Gaze-Follow [43] (GF), which contains more than 100k images with gaze target annotations.

The details of the six datasets are presented in the supplementary materials. Nevertheless, as authors have been using many subsets of G360 for evaluation, we clarify the test splits to avoid any confusion. We followed the split of [28]: G360 Full corresponds to "All 360°" (all the test set); G360 180 corresponds to "Front 180°" (gaze within 90°); and G360 40 to "Front Facing" (gaze within 20°). Additionally, we consider G360 Back (gaze above 90°) [8] and G360 Face (all detected faces), used in many studies [1, 7, 9– 11, 18, 55, 61]. G360 Face 180 or 40 corresponds to the detected face with a gaze within 90° or 20°.

4.2. Implementation Details

Each dataset has different head bounding boxes ground truth. To avoid discrepancies in cropping, we standardize the input by using a robust pre-trained head detector[2] train on the CrowdHuman dataset [47]. We match the detected and ground truth bounding boxes to get the final head crop bounding box. Furthermore, we downscale the head bounding box by 10% and resize it to 224×224 pixels. We show the impact of the head bounding box size in the supplementary materials. An 8-frame head crop clip is used for the video modalities, and the frame rate is unified across all video datasets. All the backbones used in this work are pretrained on Imagnet for static backbones and ImageNet-1K, Kinetics-400, and SUN RGB-D for Swin3D. Please refer to the supplementary materials for training and data augmentation details.

4.3. Ablation Study

Does gaze following label improve 3D gaze estimation? In Tab. 1, we evaluate various methods for learning from 2D gaze following labels (GF). We find that, with few exceptions, incorporating GF consistently improves 3D gaze estimation. This underscores the importance of broader training diversity for robust 3D gaze estimation. The specific details and advantages of each approach are discussed in the following section.

Self-Training Weakly-Supervised learning framework. In Tab. 1, we perform ablation experiments related to the

---

[2]https://github.com/zhangda1018/yolov5-crowdhuman

learning framework. In our experiments, we train with our model GaT on two 3D gaze datasets namely G360 and GFIE. The first baseline experiment is to train on a 3D gaze dataset in a standard supervised manner. Then, there are three possibilities to leverage additional 2D datasets such as GF. The first method, Weakly-Supervised (WS), applies a specific loss only to the x, y coordinates of the 3D gaze prediction for GF batch samples supervised by the 2D ground truth label. The second approach Self-Training (ST) is similar to our ST-WSGE approach described in Fig. 2 but in the second stage, 3DPred is used to supervise the training. Finally, the last approach ST-WSGE is our proposed approach described in Fig. 2 and Sec. 3.1. Compared to ST, ST-WSGE achieves higher accuracy across all evaluations except when trained on G360 and tested on EDIAP. This indicates that relying solely on 3DPred lacks diversity in gaze distribution, as it mostly follows the training data distribution. By incorporating our 3DGP label, we obtain a more robust gaze vector that enhances accuracy. Furthermore, when compared to using only 2D labels in the WS method, ST-WSGE performs slightly better overall, particularly on the frontal EDIAP and MPII benchmarks. This suggests that datasets, where the z component is significant (e.g., frontal views), require more than 2D supervision. Overall, our method is either the best or the second best (by a small difference) demonstrating the effectiveness of our ST-WSGE learning framework.

Gaze Model Network. In Tab. 2, we present the results concerning the model architecture. In our experiments, we train on G360 in a supervised manner with different training modalities combinations (I, V, and I&V). We compare our model GaT with baseline models (see Sec. 3.2) that can also handle different training data modalities. First, we notice that when models are trained on both image and video, our GaT model is the best model on G360. It suggests that spatiotemporal learning from the input is beneficial for gaze estimation, especially in non-frontal scenarios. Additionally, within each model, training on both modalities improves image evaluation but slightly reduces video evaluation. However, modality-specific models are limited to their own data type which limits the available training dataset. Indeed, in the next section, we show that combining modalities can result in cross-modal generalization. Given these findings, our model stands out as a reliable and versatile option for robust 3D gaze estimation.

Does training on additional image datasets help video generalization? As observed in Tab. 1, training our GaT model with ST-WSGE, which includes a diverse image dataset (GF), not only improves image generalization but, more notably, enhances generalization on videos. Compared to a supervised method, our approach improves image GFIE evaluation by 38% and video GFIE by 34%. A similar trend is observed when training on GFIE and evalu-

ating on G360 Full. Furthermore, the modality of the training data plays an important role in generalization, as observed in Tab. 3. By considering G360 both as an image and video dataset (I&V setting), G360I&V+GF further improves video generalization compared to G360V+GF with an angular error reduced by 2.2°. A Similar trend is observed when training on GFIE. Interestingly, adding image training data improves video generalization. Note that image-only training G360I+GF has reduced video generalization.

Temporal Context. Temporal dynamic plays a crucial role in unconstrained gaze estimation, as evidenced in Tab. 1 with our GaT model trained on both modalities (I&V). In all configurations, video predictions consistently outperform image-based predictions. Other important observations emerged from visual and quantitative analyses and are discussed in the supplementary material.

4.4. Comparison with State-of-the-art (SOTA).

Within-Dataset Experiments. In the following, we focus on within-dataset experiments. In Tab. 4, we compare our results with the state-of-the-art methods on G360 and GFIE. We report results using our model GaT in a supervised manner and using our ST-WSGE learning framework. Our model trained in a supervised manner is SOTA on image G360 Full and GFIE with 2% and 13% relative improvement, respectively. On video inference, MSA+Seq is slightly better (12.6° vs 12.5° ours) since it uses an average of multiple input scales. More importantly, when trained with gaze following labels like GF using our ST-WSGE learning framework, we outperform all the SOTA on image and video by 5% on G360 Full image, 3% on G360 Full video, and 36% on GFIE image. In contrast, Kothari *et al*. [31] don't improve when using the LAEO label (AVA) in a weakly-supervised framework. Additionally, in Tab. 5, we compared our method trained on G360 Full to methods trained on detected face subset G360 Face. Given that the state-of-the-art methods are specifically designed for near-frontal faces, our supervised model GaT is not SOTA but demonstrates competitive performance. When including gaze following label using our ST-WSGE framework, it shows very competitive results and SOTA performance on G360 Face video (9.92° vs 10.05°), G360 Face 180 video (9.84° vs 9.75°), and G360 Face 40 (8.62° vs 8.30°). Therefore, compared to methods using tight face crops (increasing eye resolution), our ST-WSGE approach proved to be competitive on near frontal view. We include a comparison with SOTA trained on G360 and evaluated on constrained benchmarks MPII and EDIAP in supplementary materials.

Cross-Dataset Experiments. In this section, we emphasize on cross dataset experiments. In Tab. 4, we compare our method with SOTA methods on generalization on G360 (bottom part). Among the few approaches that explore gen-

Method	Training Data	G360 Full		G360 180		G360 40		G360 Face		G360 Back		GFIE	
		Img	Vid	Img	Vid	Img	Vid	Img	Vid	Img	Vid	Img	Vid
Gaze360 [28]	G360I	15.6	-	13.4	-	13.2	-	-	-	-	-	-	-
Kothari et al. [31]	G360I	15.07	-	-	-	10.94	-	-	-	-	-	-	-
MSA [8]	G360I	13.9	-	12.2	-	-	-	-	-	23.5	-	-	-
Gaze360 [28]	G360V	-	13.5	-	11.4	-	11.1	-	-	-	-	-	-
Kothari et al. [31]	G360V	-	13.2	-	-	-	10.1	-	-	-	-	-	-
Kothari et al. [31]	G360V+AVA	-	13.2	-	-	-	10.2	-	-	-	-	-	-
MCGaze [22]	G360V	-	12.96	-	10.74	-	10.02	-	-	-	-	-	-
MSA+Seq [8]	G360V	-	12.5	-	10.7	-	-	-	-	-	19.0	-	-
Supervised (GaT)	G360I&V	13.64	12.60	11.66	10.67	9.10	8.23	11.20	10.25	20.74	19.53	21.86	20.89
ST-WSGE (GaT)	G360I&V+GF	**13.19**	**12.17**	**11.34**	**10.35**	**8.58**	**7.67**	**10.84**	**9.92**	**19.82**	**18.72**	**15.90**	**15.51**
GFIE [25]	GFIEI	-	-	-	-	-	-	-	-	-	-	17.7	-
Kothari et al. [31]	ETH	52.6	-	-	-	20.5	-	-	-	-	-	-	-
Kothari et al. [31]	ETH+AVA	-	25.0	-	-	-	16.9	-	-	-	-	-	-
3DGazeNet [53]†	ETH	-	-	-	-	-	-	22.1†	-	-	-	-	-
3DGazeNet [53]†	ETH+AVA+CMU	-	-	-	-	-	-	17†	-	-	-	-	-
3DGazeNet [53]†	ETH+ITWG-MV	-	-	-	-	-	-	15.4†	-	-	-	-	-
Supervised (GaT)	GFIEI&V	30.57	29.90	29.08	28.65	33.43	32.94	28.87	28.56	35.95	34.40	15.70	15.44
ST-WSGE (GaT)	GFIEI&V+GF	**21.48**	21.06	20.61	20.32	26.55	26.73	20.46	20.23	**24.61**	**23.75**	12.99	12.68
Supervised (GaT)	GFIEI&V+MPS	25.75	24.29	20.35	19.42	16.35	15.49	19.07	18.28	45.19	41.81	15.61	15.38
ST-WSGE (GaT)	GFIEI&V+MPS+GF	21.59	**20.02**	**17.02**	**15.67**	**13.90**	**12.66**	16.00	14.75	38.02	35.69	**12.82**	**12.49**

Table 4. **Comparison with the state-of-the-art on physically unconstrained benchmark Gaze360 and GFIE test set.** We report both within and cross-dataset evaluation trained using GaT model with and without our ST-WSGE framework. The top table presents methods trained with Gaze360, while at the bottom, methods are trained using GFIE. The method with † is restricted with only frontal pose with face and eye crop as input. Moreover, the method is evaluated on a new subset (head pose yaw $\in$ [-90,90]), which is close to, but not the same as, the Gaze360 Face subset.

Method	Training Data	G360 Face		G360 Face 180		G360 Face 40	
		Img	Vid	Img	Vid	Img	Vid
FullFace [61]	G360I Face	14.99	-	-	-	-	-
Dilated [9]	G360I Face	13.73	-	-	-	-	-
RT-Gene [18]	G360I Face	12.26	-	-	-	-	-
CA-Net [11]	G360I Face	11.20	-	-	-	-	-
Gaze360 [12]	G360V Face	-	11.04	-	-	-	-
ResNet50 [10]	G360I Face	10.73	-	-	-	-	-
GazeTR [10]	G360I Face	**10.62**	-	-	-	-	-
L2CS [1]	G360I Face	-	-	10.4	-	9.0	-
SPMCCA [55]	G360I Face	-	-	**10.16**	-	8.62	-
SAM-LSTM [26]	G360V Face	-	10.05	-	9.84	-	**6.92**
Supervised (GaT)	G360I&V	11.20	10.25	11.01	10.09	8.81	8.02
ST-WSGE (GaT)	G360I&V+GF	10.84	**9.92**	10.65	**9.75**	**8.30**	7.48

Table 5. **Comparison with the state-of-the-art constrained methods tested on Gaze360 detected face.** All the state-of-the-art methods use a face crop as input and are trained on the detected face subset of Gaze360. We report results trained on G360 Full using GaT and with or without ST-WSGE. Note that the other methods are constrained to face input therefore our method is more general and can be applied to any head pose orientation.

eralization on Gaze360, Kothari et al. [31] provides the most relevant comparison. In contrast, 3DGazeNet [53] provides cross-dataset generalization on G360 Face but only works on frontal faces requiring face and eye crop as input. Our results show that our ST-WSGE framework trained with various available 3D gaze datasets (GFIE, or GFIE+MPS) always improves generalization. For instance, when tested on G360 Full image and video, it always outperforms our supervised approach by 40% and 20% when trained with GFIE or GFIE+MPS, respectively. A similar trend is observed when trained on G360 and tested on GFIE. Therefore, it confirms that our framework using gaze follow labels is effective for improved generalization.

When compared to Kothari et al. [31] trained on LAEO labels (AVA), our ST-WSGE approach trained using GFIE+GF shows better performance on G360 Full but is behind on G360 40 because GFIE doesn't contain frontal samples. In contrast, when trained using GFIE+MPS+GF, it improves over Kothari on both G360 Full and 40.

Limitations. As expected when using our framework, the generalization improvement is tight to the training diversity used in the pre-training stage. In cross-dataset experiments in Tab. 4, compared to our supervised model, we observe that when trained using GFIE our framework improves more on G360 Back and less on G360 40 because of the non-frontal distribution of GFIE.

5. Conclusion

In this work, we introduced the ST-WSGE learning framework, which leverages weakly annotated images with 2D gaze datasets, such as gaze-follow labels, to enhance appearance diversity and broaden gaze distributions across natural scenes. We also presented our Gaze Transformer, GaT, which improves performance and supports both image and video training. By combining ST-WSGE and GaT, we achieve significant gains in both within- and cross-dataset experiments, reaching state-of-the-art results on GFIE and Gaze360. Additionally, we demonstrate effective cross-modal generalization, a critical capability given the scarcity of video datasets. We believe our approach is a promising solution for robust 3D gaze tracking in the wild, suitable for a range of challenging applications.

References

[1] Ahmed A Abdelrahman, Thorsten Hempel, Aly Khalifa, Ayoub Al-Hamadi, and Laslo Dinges. L2cs-net: Fine-grained gaze estimation in unconstrained environments. In *2023 8th International Conference on Frontiers of Signal Processing (ICFSP)*, pages 98–102. IEEE, 2023. 2, 4, 6, 8, 1

[2] Henny Admoni and Brian Scassellati. Social eye gaze in human-robot interaction: a review. *Journal of Human-Robot Interaction*, 6(1):25–63, 2017. 2

[3] Sean Andrist, Xiang Zhi Tan, Michael Gleicher, and Bilge Mutlu. Conversational gaze aversion for humanlike robots. In *Proceedings of the 2014 ACM/IEEE international conference on Human-robot interaction*, pages 25–32, 2014. 1

[4] Yiwei Bao, Yunfei Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with rotation consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4207–4216, 2022. 3, 2

[5] Alisa Burova, John Mäkelä, Jaakko Hakulinen, Tuuli Keskinen, Hanna Heinonen, Sanni Siltanen, and Markku Turunen. Utilizing vr and gaze tracking to develop ar solutions for industrial maintenance. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13, 2020. 1

[6] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 2

[7] Andy Catruna, Adrian Cosma, and Emilian Radoi. Crossgaze: A strong method for 3d gaze estimation in the wild. In *International Conference on Automatic Face and Gesture Recognition (FG)*, 2024. 2, 6, 1

[8] Chu-Song Chen, Hsuan-Tien Lin, et al. 360-degree gaze estimation in the wild using multiple zoom scales. In *British Machine Vision Conference (BMVC)*, 2021. 3, 4, 6, 8, 1, 2

[9] Zhaokang Chen and Bertram E Shi. Appearance-based gaze estimation using dilated-convolutions. In *Asian Conference on Computer Vision*, pages 309–324. Springer, 2018. 2, 3, 6, 8, 1

[10] Yihua Cheng and Feng Lu. Gaze estimation using transformer. In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 3341–3347. IEEE, 2022. 3, 4, 8

[11] Yihua Cheng, Shiyao Huang, Fei Wang, Chen Qian, and Feng Lu. A coarse-to-fine adaptive network for appearance-based gaze estimation. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10623–10630, 2020. 2, 6, 8, 1

[12] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based gaze estimation with deep learning: A review and benchmark. *arXiv preprint arXiv:2104.12668*, 2021. 8, 1, 2

[13] Yihua Cheng, Yiwei Bao, and Feng Lu. Puregaze: Purifying gaze feature for generalizable gaze estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 436–443, 2022. 3, 2

[14] Eunji Chong, Nataniel Ruiz, Yongxin Wang, Yun Zhang, Agata Rozga, and James M. Rehg. Connecting gaze, scene, and attention: Generalized attention estimation via joint modeling of gaze and scene saliency. In *The European Conference on Computer Vision (ECCV)*, 2018. 2

[15] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M. Rehg. Detecting attended visual targets in video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 4, 6, 7

[16] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 4, 5

[17] Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022. 5

[18] Tobias Fischer, Hyung Jin Chang, and Yiannis Demiris. Rt-gene: Real-time eye gaze estimation in natural environments. In *Proceedings of the European conference on computer vision (ECCV)*, pages 334–352, 2018. 2, 3, 6, 8, 1

[19] 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 *Proceedings of the symposium on eye tracking research and applications*, pages 255–258, 2014. 2, 3, 6, 1

[20] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens Van Der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16102–16112, 2022. 2, 5

[21] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnimae: Single model masked pretraining on images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10406–10417, 2023. 2, 5

[22] Yiran Guan, Zhuoguang Chen, Wenzheng Zeng, Zhiguo Cao, and Yang Xiao. End-to-end video gaze estimation via capturing head-face-eye spatial-temporal interaction context. *IEEE Signal Processing Letters*, 30:1687–1691, 2023. 3, 8

[23] Anshul Guptat, Samy Tafasca, Arya Farkhondeh, Pierre Vuillecard, and Jean-Marc Odobez. Mtgs: A novel framework for multi-person temporal gaze following and social gaze prediction. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 3

[24] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 1439–1449, 2021. 5

[25] Zhengxi Hu, Yuxue Yang, Xiaolin Zhai, Dingye Yang, Bohan Zhou, and Jingtai Liu. Gfie: A dataset and baseline for gaze-following from 2d to 3d in indoor environments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8907–8916, 2023. 2, 3, 6, 8, 1

[26] Swati Jindal, Mohit Yadav, and Roberto Manduchi. Spatio-temporal attention and gaussian processes for personalized

video gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 604–614, 2024. 3, 4, 8

[27] Isaac Kasahara, Simon Stent, and Hyun Soo Park. Look both ways: Self-supervising driver gaze estimation and road scene saliency. In *European Conference on Computer Vision*, pages 126–142. Springer, 2022. 1

[28] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically unconstrained gaze estimation in the wild. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6912–6921, 2019. 2, 3, 4, 6, 8, 1

[29] Chris L Kleinke. Gaze and eye contact: a research review. *Psychological bulletin*, 100(1):78, 1986. 2

[30] D. Klotz, J.Wienke, B. Wrede, S. Wrede, S. Sheikhi, D. Jayagopi, V. Khalidov, and J.-M. Odobez. Robot-to-group interaction in a vernissage: Architecture & dataset for multi-party dialog. In *Proceedings of 5th International Conference on Cognitive Systems*, 2012. 2

[31] Rakshit Kothari, Shalini De Mello, Umar Iqbal, Wonmin Byeon, Seonwook Park, and Jan Kautz. Weakly-supervised physically unconstrained gaze estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9980–9989, 2021. 2, 3, 4, 5, 7, 8

[32] Kyle Krafka, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba. Eye tracking for everyone. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2176–2184, 2016. 2

[33] Gang Liu, Yu Yu, Kenneth Alberto Funes Mora, and Jean-Marc Odobez. A differential approach for gaze estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI),*, 43(3):1092–1099, 2021. 3

[34] Yunfei Liu, Ruicong Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with outlier-guided collaborative adaptation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3835–3844, 2021. 2

[35] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021. 4, 5

[36] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 5

[37] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *In International Conference on Learning Representations*, 2018. 2

[38] Manuel J Marin-Jimenez, Vicky Kalogeiton, Pablo Medina-Suarez, and Andrew Zisserman. Laeo-net: revisiting people looking at each other in videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3477–3485, 2019. 2, 5

[39] Michael McCloskey and Neal J Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. In *Psychology of learning and motivation*, pages 109–165. Elsevier, 1989. 2

[40] Soma Nonaka, Shohei Nobuhara, and Ko Nishino. Dynamic 3d gaze from afar: Deep gaze estimation from temporal eye-head-body coordination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2192–2201, 2022. 2, 3

[41] Cristina Palmero, Javier Selva, Mohammad Ali Bagheri, and Sergio Escalera. Recurrent cnn for 3d gaze estimation using appearance and shape cues. In *In British Machine Vision Conference (BMVC)*, 2018. 3, 4

[42] Seonwook Park, Emre Aksan, Xucong Zhang, and Otmar Hilliges. Towards end-to-end video-based eye-tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 747–763. Springer, 2020. 2, 3, 4

[43] Adria Recasens, Aditya Khosla, Carl Vondrick, and Antonio Torralba. Where are they looking? *Advances in neural information processing systems*, 28, 2015. 2, 6, 1

[44] R. Reiter-Palmon, T. Sinha, J. Gevers, J.-M. Odobez, and G. Volpe. Theories and models of teams and group. *Journal of Small Group Research*, 45(5):544–567, 2017. 1

[45] Alexander Richard, Colin Lea, Shugao Ma, Jurgen Gall, Fernando De la Torre, and Yaser Sheikh. Audio-and gaze-driven facial animation of codec avatars. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 41–50, 2021. 1

[46] Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, et al. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *International Conference on Machine Learning*, pages 29441–29454. PMLR, 2023. 4

[47] Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. Crowdhuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018. 6

[48] S. Sheikhi and J.M. Odobez. Combining dynamic head pose and gaze mapping with the robot conversational state for attention recognition in human-robot interactions. *Pattern Recognition Letters*, 66:81–90, 2015. 2

[49] R. Siegfried and J.-M. Odobez. Robust unsupervised gaze calibration using conversation and manipulation attention priors. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(1):1–27, 2022. 3

[50] S. Tafasca, A. Gupta, and J.-M. Odobez. Childplay: A new benchmark for understanding children's gaze behaviour. In *Int. Conference on Computer Vision (ICCV), Paris*, 2023. 2

[51] Samy Tafasca, Anshul Guptat, and Jean-Marc Odobez. Toward semantic gaze target detection. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 2

[52] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 5

[53] Evangelos Ververas, Polydefkis Gkagkos, Jiankang Deng, Michail Christos Doukas, Jia Guo, and Stefanos Zafeiriou. 3dgazenet: Generalizing gaze estimation with weak-supervision from synthetic views. In *ECCV*, 2024. 3, 8, 2

[54] Yaoming Wang, Yangzhou Jiang, Jin Li, Bingbing Ni, Wenrui Dai, Chenglin Li, Hongkai Xiong, and Teng Li. Contrastive regression for domain adaptation on gaze estimation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19376–19385, 2022. 3, 1, 2

[55] Chao Yan, Weiguo Pan, Cheng Xu, Songyin Dai, and Xuewei Li. Gaze estimation via strip pooling and multi-criss-cross attention networks. *Applied Sciences*, 13(10): 5901, 2023. 2, 4, 6, 8, 1

[56] Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. Wider face: A face detection benchmark. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5525–5533, 2016. 3, 1

[57] Yu Yu and Jean-Marc Odobez. Unsupervised representation learning for gaze estimation. In *Int. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 3

[58] Yu Yu, Gang Liu, and Jean-Marc Odobez. Improving few-shot user-specific gaze adaptation via gaze redirection synthesis. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3

[59] Mingfang Zhang, Yunfei Liu, and Feng Lu. Gazeonce: Real-time multi-person gaze estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4197–4206, 2022. 3, 6, 1

[60] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4511–4520, 2015. 2, 3, 6, 1

[61] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas 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*, pages 51–60, 2017. 2, 3, 6, 8, 1

[62] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Revisiting data normalization for appearance-based gaze estimation. In *Proceedings of the 2018 ACM symposium on eye tracking research & applications*, pages 1–9, 2018. 2

[63] 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 *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 365–381. Springer, 2020. 2, 1