

# UniGaze: Towards Universal Gaze Estimation via Large-scale Pre-Training

## Supplementary Material

Training Data \ Test	$X_{\text{test}}$	$M_{\text{test}}$	$GC_{\text{test}}$	$E_{\text{test}}$	$G360_{\text{test}}$
ResNet-50					
<i>same-domain</i>	5.25	5.11	3.49	8.51	11.87
<i>leave-one-dataset-out</i>	16.31 ( $\uparrow 210.7\%$ )	6.23 ( $\uparrow 21.9\%$ )	6.35 ( $\uparrow 82.0\%$ )	8.25 ( $\downarrow 3.1\%$ )	20.38 ( $\uparrow 71.7\%$ )
<i>joint-dataset</i>	<b>5.04</b> ( $\downarrow 4.0\%$ )	5.88 ( $\uparrow 15.1\%$ )	3.59 ( $\uparrow 2.9\%$ )	<b>6.04</b> ( $\downarrow 29.0\%$ )	<b>10.55</b> ( $\downarrow 11.1\%$ )
UniGaze-H					
<i>same-domain</i>	4.62	5.19	3.01	6.11	9.44
<i>leave-one-dataset-out</i>	11.29 ( $\uparrow 144.4\%$ )	5.22 ( $\uparrow 0.6\%$ )	5.13 ( $\uparrow 70.4\%$ )	6.14 ( $\uparrow 0.5\%$ )	13.12 ( $\uparrow 39.0\%$ )
<i>joint-dataset</i>	<b>4.46</b> ( $\downarrow 3.5\%$ )	<b>5.08</b> ( $\downarrow 2.1\%$ )	3.20 ( $\uparrow 6.3\%$ )	<b>5.16</b> ( $\downarrow 15.6\%$ )	<b>9.07</b> ( $\downarrow 3.9\%$ )

Table 1. Comparison of different training data configurations for gaze estimation. Each column represents a specific test dataset: XGaze Test, MPIIFaceGaze Test, GazeCapture Test, EYEDIAP Test, and Gaze360 Test. Each row corresponds to a training configuration: *Same-domain* means training on the same domain as the test set, *leave-one-dataset-out* means training on the remaining four datasets other than the test set, and *joint-dataset* means training on the aggregated Train split of all five datasets. The percentages in parentheses indicate the reduction or increment compared to the *same-domain* results, where lower errors indicate better performance. For the *leave-one-dataset-out* configuration, the errors reported here are on the Test splits, while the main paper reports errors on the entire dataset.

In this supplementary material, we first provide an analysis of the effect of combining multiple domains. Then, we include additional ablation studies to investigate the effects of color-jitter augmentation and pixel normalization during the MAE pre-training. Finally, we present qualitative results, highlighting images captured under diverse and challenging conditions.

### 1. Analysis on Combining Multiple Domains

We analyze the effect of different training data configurations on gaze estimation performance. Specifically, we compare three configurations: training on the same domain (*same-domain*), training on multiple domains excluding the testing domain (*leave-one-dataset-out*), and training on multiple domains including the testing domain (*joint-dataset*).

Table 1 shows the comparison of gaze errors for these configurations. Each column corresponds to a specific test dataset: XGaze Test, MPIIFaceGaze Test, GazeCapture Test, EYEDIAP Test, and Gaze360 Test, while each row represents a training configuration. This *same-domain* setting is different from the *within-dataset* in the main paper. We use the splits defined in Sec. 4.1 of the main paper. Especially, please note that for MPIIFaceGaze dataset, we train the model on the first 10 subjects and test on the remaining five subjects, different from the typical leave-one-subject-out protocol [1, 14, 16].

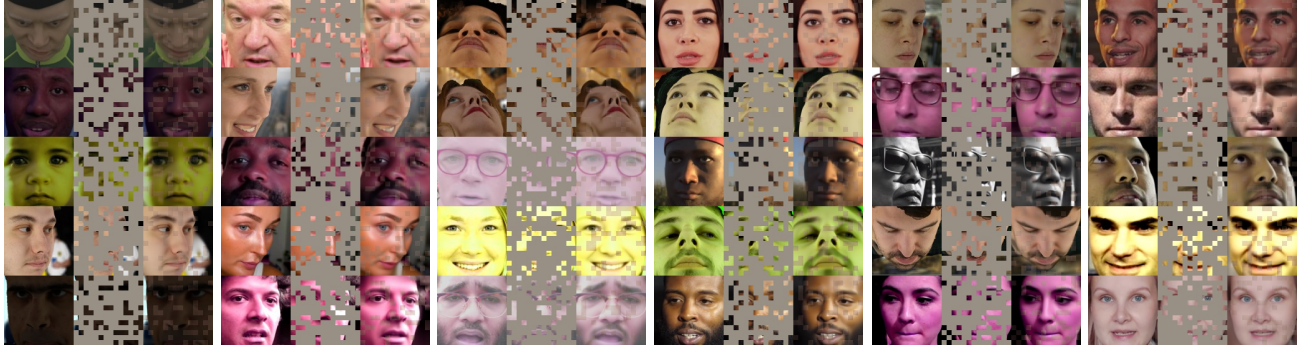
The percentages in parentheses indicate the reduction or increment compared to the *same-domain* results, where

lower errors indicate better performance. Note that, for the *leave-one-dataset-out* configuration, errors on the entire left-out dataset are reported in the main paper, but here we present errors on the Test split to align with the other configurations that require dataset splits.

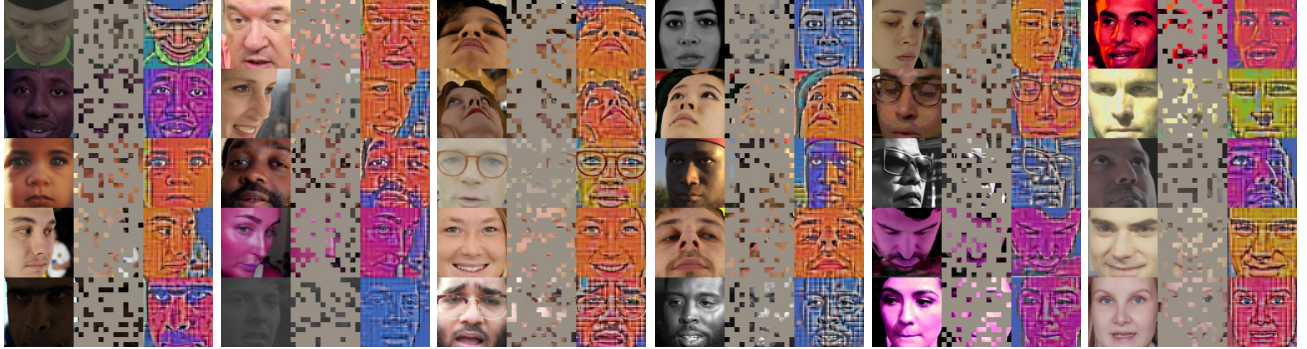
**Same-domain** In general, training and testing on the same domain (*same-domain*) yields the best results, even though datasets combined from multiple domains have the potential to be more diverse. This emphasizes the persistent challenge of achieving optimal performance when using data from different domains. The exception observed for  $E_{\text{test}}$  with the ResNet-50 backbone may be attributed to the limited number of samples in the EYEDIAP Train split.

**Leave-one-dataset-out** In the *leave-one-dataset-out* configuration, we observe varying tendencies across different test datasets. Some datasets achieve errors comparable to the *same-domain* results, while others remain challenging. For instance, for  $M_{\text{test}}$  and  $E_{\text{test}}$ , which are relatively less complex, the remaining four datasets provide sufficient information to achieve good performance. In contrast, for  $X_{\text{test}}$ ,  $GC_{\text{test}}$ , and  $G360_{\text{test}}$ , the remaining four datasets fail to fully capture the critical factors required for optimal performance. This variation highlights the strong dependence of performance on the attributes of the training data.

Importantly, our UniGaze-H demonstrates a smaller performance gap compared to ResNet-50 in most cases, with the only exception being EYEDIAP, where the difference is



(a) MAE reconstruction examples without pixel normalization.



(b) MAE reconstruction examples with pixel normalization (Proposed).

Figure 1. Examples comparison of the pixel normalization during the MAE pre-training. The left, middle, and right columns show the original image, masked input, and the reconstructed image, respectively.

marginal. This suggests that UniGaze-H is better equipped to learn gaze representations from out-of-domain data with less overfitting, underscoring its enhanced generalization capability.

**Joint-dataset** Overall, the *joint-dataset* configuration demonstrates significant promise, creating a single model robust across multiple test domains. For UniGaze-H, the only exception is  $GC_{test}$ , where the *joint-dataset* configuration produces a slightly higher error ( $3.01 \rightarrow 3.20$ ). Although this suggests some negative effects from the other four datasets, the effects remain marginal. While the improvement percentages for UniGaze-H are smaller compared to ResNet-50, the absolute errors are consistently lower.

## 2. Additional Ablation Studies on Pre-Training

**Effect of Pre-Training Dataset Composition** Beyond the overall pre-training dataset size, the composition of the dataset also plays a critical role in learning effective face representations. To investigate the impact of different facial dataset components, we conduct an experiment where we incrementally accumulate datasets during the MAE pre-training stage and analyze their effect on the



Figure 2. Effect of MAE pre-training dataset composition on downstream gaze estimation performance. The horizontal axis represents the incremental accumulation of datasets, while the vertical axis shows the percentage reduction in error relative to the first CelebV-Text dataset [20].

downstream gaze estimation performance. Starting with CelebV-Text [20], we progressively add datasets for pre-training and evaluate the models separately on gaze esti-

mation. Each pre-trained model is subsequently trained on gaze datasets using the same *leave-one-dataset-out* protocol. Figure 2 illustrates the error change across different test sets as more datasets are included in pre-training.

Overall, the results indicate that adding more diverse data during pre-training generally enhances gaze generalization. However, there are exceptions that adding a dataset can result in increased error for specific test sets. For example, adding VFHQ increases the error on the XGaze Test set from 12.65 to 13.32, while including SFHQ-T2I causes performance fluctuations across different benchmarks. This suggests that certain dataset attributes may not align well with particular test distributions, leading to suboptimal transferability. On the other hand, datasets such as VG-Face2 and XGaze-Dense provide performance improvements on most test sets. Additionally, performance gains becomes marginal as the dataset number increases, aligning with the analysis of pre-training data size in the main paper.

In conclusion, dataset diversity plays a crucial role in improving MAE pre-training for gaze estimation. A more detailed analysis of dataset attributes and their impact on gaze estimation remains an open research question, which we leave for future work. Nonetheless, our empirical results suggest that increasing data diversity in pre-training tends to improve model performance across various test domains.

### Effect of Novel-View Synthesis Data in Pre-Training

To examine the effect of novel-view synthesis in pre-training data, we conduct further experiments separating these two elements. In Tab. 2, we conduct an ablation study by varying data subsets during the pre-training: real datasets (CelebV-Text, VGGFace2, and VFHQ), synthetic datasets (FaceSynthetics and SFHQ-T2I), and novel-view-rendered datasets (FFHQ-NV and XGaze-Dense). We use the UniGaze-B to conduct the experiment due to its time efficiency. After pre-training, we train on XGaze and test on the rest of the four datasets.

The results further clarify the effect of different data types on the model’s generalizability. Adding synthetic data (*Real* + *Syn.*) reduces errors in several test domains compared to using only real data, suggesting the variability of the synthetic data contributes to generalization. Further incorporating novel-view data (*Real* + *Syn.* + *NV*) provides additional performance gains, especially in head-pose generalization, likely due to the expanded range of facial orientations. This finding supports the idea that a mix of real, synthetic, and novel-view data in MAE pre-training strengthens ViT’s representation learning.

**Effect of Pixel Normalization** The patch normalization technique is applied during the MAE pre-training as suggested in [7] which is different from reconstructing the natural image, as shown in Fig. 1. We compare models pre-

<i>Real</i>	<i>Syn.</i>	<i>NV.</i>	<b>M</b>	<b>GC</b>	<b>E</b>	<b>G360</b>
✓			6.79	7.81	6.86	12.93
✓	✓		6.57	7.37	<b>6.51</b>	13.23
✓	✓	✓	<b>6.21</b>	<b>7.35</b>	6.64	<b>12.18</b>

Table 2. We ablate the pre-training facial datasets by comparing real, synthetic, and novel-rendered images. The comparison is performed on the UniGaze-B network, followed by training on XGaze. The last row represents the full-dataset setting.

<b>Color-Jitter</b>	<b>Pixel Norm.</b>	<b>M</b>	<b>GC</b>	<b>E</b>	<b>G360</b>
✗	✗	7.52	8.01	8.56	14.14
✓	✗	7.17	8.23	8.03	14.03
✗	✓	7.18	7.94	8.05	13.66
✓	✓	<b>6.21</b>	<b>7.35</b>	<b>6.64</b>	<b>12.18</b>

Table 3. Ablation studies on the pre-training, comparing the effect of the color-jitter augmentation and the pixel normalization. During the gaze estimation training, we train the model using XGaze and test on the other four datasets to evaluate the generalizability.

trained with and without patch normalization to investigate its impact.

**Effect of Color-Jitter Augmentation** Color jittering introduces randomness in brightness, contrast, saturation, and hue to simulate diverse lighting conditions, enhancing the robustness of learned features. We compare models pre-trained with and without color-jitter augmentation to investigate its impact.

**Results** We use the UniGaze-B model as the backbone and compare different pre-training settings, followed by training on XGaze and testing on the remaining four datasets. Table 3 demonstrates that both color-jitter augmentation and pixel normalization contribute to improved gaze estimation performance, highlighting their benefits for the generalization of the pre-trained model. Notably, pixel normalization consistently improves performance across all test datasets, aligning with the observations in the original MAE paper [7], which showed that pixel normalization enhances the following gaze estimation training.

## 3. Few-Shot Personalization

We evaluate the few-shot performance to assess the model’s ability to adapt to new subjects with minimal data and less overfitting. For this evaluation, we use the models trained in the *leave-one-dataset-out* evaluation and fine-tune the model on a small number of subjects from the remaining dataset. For each subject, we freeze the ViT backbone and fine-tune only the MLP head using  $K \in \{1, 2, 4, 8, 16, 32\}$  samples.



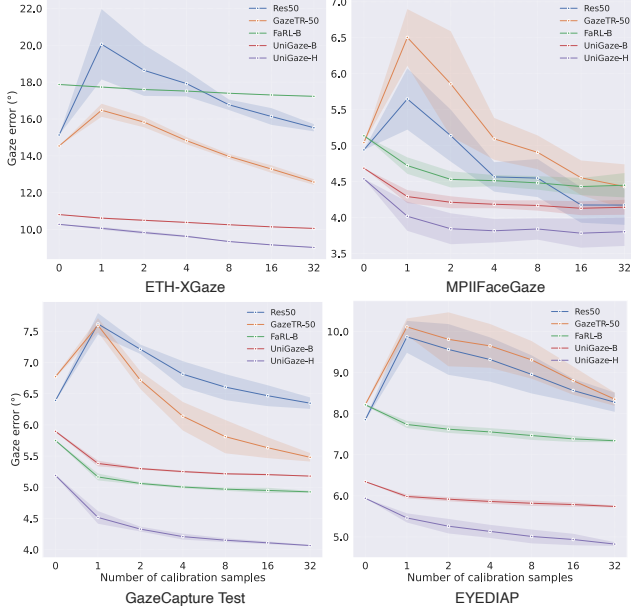


Figure 3. Few-shot personalization on ETH-XGaze, MPIIFaceGaze, GazeCapture<sub>Test</sub>, and EYEDIAP. The vertical axis means the gaze error (°), and the horizontal axis represents the number of calibration samples  $K$ .

For ETH-XGaze, GazeCapture<sub>Test</sub>, and EYEDIAP, we define the calibration pool as the first 128 images of each subject and use all remaining images for evaluation. For MPIIFaceGaze, following [22], we use the first 2,500 images as the calibration pool and the last 500 images for evaluation. Subjects with fewer than 500 images are excluded from this analysis.

We repeat five trials with different random draws of the  $K$  calibration samples. Figure 3 reports the mean and standard deviation across these trials. We can observe that UniGaze adapts effectively even with a single labeled sample ( $K=1$ ) whereas CNN baselines show limited gains. As  $K$  increases, UniGaze yields steady gains with a clear margin over other baselines.

#### 4. Comparison with the SOTAs

3DGazeNet [16] collects in-the-wild face images with pseudo gaze labels and applies multi-view synthesis to obtain an augmented dataset ITWG-MV. To account for the difference in test data settings, we compare 3DGazeNet with UniGaze-H separately in Tab. 4. The results demonstrate that UniGaze-H outperforms 3DGazeNet in all domain generalization settings.

**Re-implementation** In the main paper, we compared our UniGaze-H model with state-of-the-art (SOTA) methods

Models	$X \rightarrow M$	$X \rightarrow GC$	$G360 \rightarrow M$	$G360 \rightarrow GC$
3DGazeNet <sup>†</sup> [16]	6.0	7.8	6.3	8.0
UniGaze-H	<b>5.57</b>	<b>6.56</b>	<b>5.65</b>	<b>7.23</b>

Table 4. Domain generalization compared with SOTA methods. The results marked with <sup>†</sup> are directly cited from previous studies [16].

Models	$X \rightarrow M$	$X \rightarrow E_{CS}$	$G360 \rightarrow M$	$G360 \rightarrow E_{CS}$
ResNet-18 <sup>†</sup> [24]	8.02	9.11	8.04	9.20
ResNet-18	7.57	9.54	9.24	8.07
PureGaze <sup>†</sup> [5]	7.08	7.48	9.28	9.32
PureGaze	6.68	7.62	8.87	10.53
UniGaze-H	<b>5.57</b>	<b>4.65</b>	<b>5.65</b>	<b>5.86</b>

Table 5. Domain generalization compared with SOTA methods and their re-implementations. The results marked with <sup>†</sup> are cited from previous studies [5, 24], and the rest of the results are based on our implementation.

Dataset	Type	# Identities	# Samples
CelebV-Text [20]	Real	13,179	666,967
VFHQ [18]	Real	10,382	231,809
VGGFace2 [4]	Real	9,131	182,603
FaceSynthetics [17]	Syn.	86,878 <sup>†</sup>	86,878
SFHQ-T2I [2]	Syn.	120,241 <sup>†</sup>	120,241
FFHQ-NV [8, 12]	Syn.	25,000	100,000
XGaze-Dense [11, 13, 23]	Syn.	60	267,160
Total	-	264,871	1,655,668

Table 6. Statistics of face datasets used to pre-train UniGaze in terms of data type to be real or synthetic (Syn.), number of identities, and number of samples. The <sup>†</sup> indicates that we assume there are no duplicated identities during the synthesis image generation.

using their reported results. It is important to note that minor discrepancies may arise due to differences in our data pre-processing compared to prior works [5, 19, 24]. To ensure a fair comparison, we re-implemented ResNet-18 and PureGaze [5] using our pre-processed datasets, aligning them with the reported results [5, 24]. The re-implementation results, alongside the reported values, are summarized in Tab. 5.

While minor differences exist between our re-implementation and the reported values, the improvements achieved by our UniGaze-H model remain significant, demonstrating its superior performance across all domain generalization tasks.



## 5. Implementation Details

**Pre-Training Datasets** Table 6 summarizes our pre-training dataset composition, which shows that our pre-training data covers both real and synthetic, and a large number of identities and samples.

**Novel-Rendered Data Preparation** To render images from novel views, we follow the rendering approach described in [12]. To control the head pose, we randomly generate target head poses and compute the corresponding rotation matrices to apply to the 3D face models. During the rendering process, 40% of the images are assigned a random background color, while the remaining 60% use random scene images from the Places365 dataset [25] as background. Additionally, to simulate varied lighting conditions, half of the rendered images are adjusted to have lower ambient light intensity, ranging from 0.2 to 0.75.

All face images in our method are in the size of  $224 \times 224$  after the data normalization process [21]. When the camera parameters are unknown, we use a camera matrix with focal length  $f$  set to the image width and principal point  $(c_x, c_y)$  set to half the image height and width.

**Pre-Training** We apply random color-jitter augmentation with a probability of 0.5 and the following parameters: hue in the range  $[-0.15, 0.15]$ , saturation in  $[0.8, 1.2]$ , contrast in  $[0.4, 1.8]$ , and brightness in  $[0.7, 1.3]$ . We apply random grayscale with a probability of 0.05 on all images.

**Gaze Estimation Training** We use the Adam optimizer [9] with a learning rate of  $1 \times 10^{-4}$  and a weight decay of  $1 \times 10^{-6}$  for all experiments. For experiments with ResNet-50 and GazeTR-50, we set the batch size to 128 and decay the learning rate by 0.1 every five epochs, with a total of 12 epochs. For cross-dataset evaluation with UniGaze-H, we use a batch size of 128 and train the model for eight epochs with the one-cycle learning rate schedule [15]. For *leave-one-dataset-out* and *joint-dataset* evaluations, we set the batch size to 160 with 12 epochs.

## 6. Qualitative Results

We present additional qualitative results using the UniGaze-H model trained on the aggregated datasets under the *joint-dataset* setting. We employ an off-the-shelf facial landmark detector [3] to extract landmarks and perform data normalization. Gaze estimation is conducted on the normalized images, and the results are de-normalized back to the original image for visualization. For reference, we also include the normalized faces alongside the original images.

Figure 4 and Fig. 5 showcase examples from various in-the-wild videos captured under challenging conditions,

including large head poses and diverse lighting environments. Notably, we also include a synthetic example from URAvatar [10] (bottom row in Fig. 5), which generates faces with controlled viewpoints and lighting. In Fig. 6, we present more examples from the gaze-following dataset VideoAttentionTarget [6], a collection of diverse samples extracted from movies. This dataset provides annotated gaze targets, which are visualized when annotated within the image frame, as some targets may be out of frame.

These examples highlight the model’s ability to predict gaze direction accurately in unseen environments, even under extreme head poses, challenging lighting conditions, and synthetic appearances.

## 7. Ethical Considerations

Our research involves the use of existing facial and gaze datasets. In accordance with ethical guidelines, we rely on the fact that these datasets were originally collected and published following relevant ethical and data protection standards, including obtaining consent, and we do not generate or collect additional new data. Our experimental protocols involve only image content, without identifiable personal information or links to other personal data.

## 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. 1
- [2] David Beniaguev. Synthetic faces high quality - text 2 image (sfhq-t2i) dataset, 2024. 4
- [3] Adrian Bulat and Georgios Tzimiropoulos. How far are we from solving the 2d & 3d face alignment problem? (and a dataset of 230,000 3d facial landmarks). In *Proc. ICCV*, 2017. 5
- [4] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *Proc. FG*, pages 67–74. IEEE, 2018. 4
- [5] Yihua Cheng, Yiwei Bao, and Feng Lu. Puregaze: Purifying gaze feature for generalizable gaze estimation. In *Proc. AAAI*, pages 436–443, 2022. 4
- [6] Eunji Chong, Yongxin Wang, Nataniel Ruiz, and James M Rehg. Detecting attended visual targets in video. In *Proc. CVPR*, pages 5396–5406, 2020. 5, 9
- [7] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proc. CVPR*, pages 16000–16009, 2022. 3
- [8] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of StyleGAN. In *Proc. CVPR*, 2020. 4
- [9] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *Proc. ICLR*, 2015. 5



Figure 4. Qualitative results from various in-the-wild video examples. The normalized input images are displayed alongside the original image for reference.

- [10] Junxuan Li, Chen Cao, Gabriel Schwartz, Rawal Khirodkar, Christian Richardt, Tomas Simon, Yaser Sheikh, and Shunsuke Saito. Uravatar: Universal relightable gaussian codec avatars. *arXiv preprint arXiv:2410.24223*, 2024. 5, 7
- [11] Agisoft LLC. Agisoft metashape. <https://www.agisoft.com/>, 2024. 4
- [12] Jiawei Qin, Takuru Shimoyama, and Yusuke Sugano. Learning-by-novel-view-synthesis for full-face appearance-based 3d gaze estimation. In *Proc. CVPRW*, pages 4981–4991, 2022. 4, 5
- [13] Jiawei Qin, Takuru Shimoyama, Xucong Zhang, and Yusuke Sugano. Domain-adaptive full-face gaze estimation via novel-view-synthesis and feature disentanglement. *arXiv preprint arXiv:2305.16140*, 2023. 4
- [14] Yichen Shi, Feifei Zhang, Wenming Yang, Guijin Wang, and Nan Su. Agent-guided gaze estimation network by two-eye asymmetry exploration. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 2320–2326. IEEE, 2024. 1
- [15] Leslie N Smith and Nicholay Topin. Super-convergence:





Figure 5. Qualitative results of in-the-wild video and synthetic video. The normalized input images are displayed alongside the original image for reference.



Very fast training of neural networks using large learning rates. In *Artificial intelligence and machine learning for multi-domain operations applications*, pages 369–386. SPIE, 2019. [5](#)

- [16] Evangelos Ververas, Polydefkis Gkagkos, Jiankang Deng, Michail Christos Doukas, Jia Guo, and Stefanos Zafeiriou. 3dgazenet: Generalizing 3d gaze estimation with weak-supervision from synthetic views. In *ECCV*, pages 387–404. Springer, 2025. [1](#), [4](#)
- [17] Erroll Wood, Tadas Baltrušaitis, Charlie Hewitt, Sebastian Dziadzio, Thomas J Cashman, and Jamie Shotton. Fake it till you make it: face analysis in the wild using synthetic data alone. In *Proc. ICCV*, pages 3681–3691, 2021. [4](#)
- [18] Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. Vfhq: A high-quality dataset and benchmark for video face super-resolution. In *Proc. CVPR*, pages 657–666, 2022. [4](#)
- [19] Mingjie Xu, Haofei Wang, and Feng Lu. Learning a generalized gaze estimator from gaze-consistent feature. In *Proc. AAAI*, pages 3027–3035, 2023. [4](#)
- [20] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *Proc. CVPR*, pages 14805–14814, 2023. [2](#), [4](#)
- [21] Xucong Zhang, Yusuke Sugano, and Andreas Bulling. Re-visiting data normalization for appearance-based gaze estimation. In *Proc. ETRA*, 2018. [5](#)
- [22] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Mpiigaze: Real-world dataset and deep appearance-based gaze estimation. *IEEE TPAMI*, 41(1):162–175, 2019. [4](#)
- [23] 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 *Proc. ECCV*, pages 365–381. Springer, 2020. [4](#)
- [24] Ruijie Zhao, Pinyan Tang, and Sihui Luo. Improving domain generalization on gaze estimation via branch-out auxiliary regularization. *arXiv preprint arXiv:2405.01439*, 2024. [4](#)
- [25] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE TPAMI*, 2017. [5](#)



Figure 6. Qualitative results of examples from the VideoAttentionTarget dataset [6]. Gaze targets are visualized when annotated within the image frame, as some targets may be out of frame. The normalized input images are displayed alongside the original image for reference.