Learning spatio-temporal feature representations for video-based gaze estimation

Alexandre Personnic, Mihai Bâce
KU Leuven, Department of Computer Science, Group T Leuven Campus, Leuven, Belgium
{alexandre.personnic, mihai.bace}@kuleuven.be

Abstract

Video-based gaze estimation methods aim to capture the inherently temporal dynamics of human eye gaze from multiple image frames. However, since models must capture both spatial and temporal relationships, performance is limited by the feature representations within a frame but also between multiple frames. We propose the Spatio-Temporal Gaze Network (ST-Gaze), a model that combines a CNN backbone with dedicated channel attention and self-attention modules to fuse eye and face features optimally. The fused features are then treated as a spatial sequence, allowing for the capture of an intra-frame context, which is then propagated through time to model inter-frame dynamics. We evaluated our method on the EVE dataset and show that ST-Gaze achieves state-of-the-art performance both with and without person-specific adaptation. Additionally, our ablation study provides further insights into the model performance, showing that preserving and modelling intra-frame spatial context with our spatio-temporal recurrence is fundamentally superior to premature spatial pooling. As such, our results pave the way towards more robust video-based gaze estimation using commonly available cameras.

1. Introduction

Gaze behaviour is a powerful indicator of human intention and attention [34], making gaze estimation a valuable tool in domains ranging from interactive systems like HCI and gaming [17, 25, 32] to analytical applications in healthcare and education [15, 38, 41, 42]. To make this technology widely accessible, recent research has focused on appearance-based methods that leverage deep learning on commodity cameras, supplanting traditional model-based techniques [14]. The success of mapping static images to gaze directions [1, 29, 43] has naturally led to video-based estimation, where Recurrent Neural Networks (RNNs) are used to model the temporal dynamics of eye movements, further improving robustness and accuracy [8, 28, 33].

Despite significant progress in recent years, the performance of appearance-based models is often limited by two fundamental challenges. The first is the presence of unobservable, person-specific anatomical factors, such as the angle difference between the optical and visual axis of the eye, commonly referred to as the kappa angle [13]. This creates a subject-dependent offset that is difficult to model from images alone [18, 27]. To tackle this challenge, several person-specific adaptation and calibration techniques have been studied. Classical approaches often rely on an explicit calibration phase, where a user looks at one or more known points on a screen to map their gaze [13]. More recently, learning-based methods have emerged that can adapt a general model to a new person using only a few labelled or unlabelled samples [4, 23, 27]. In our work, we evaluate our model's feature representation by using it as a backbone for the Self-Calibration and Person-specific Transform (SCPT) modules proposed by Bao et al. [4]. We chose SCPT as it represents the state-of-the-art in few-shot adaptation on the EVE dataset and does not require labelled calibration data, aligning with our goal of unconstrained gaze estimation. Our experiments demonstrate that our proposed video-based gaze estimation model achieves state-of-the-art performance when these downstream adaptation techniques are used.

The second, and the focus of our work due to its broad implications for general-purpose video-based gaze estimation, is the quality of feature representations extracted from the input images. Recent work has demonstrated that incorporating temporal context from video sequences can improve accuracy [28, 33]. However, we hypothesize that the performance of the temporal models is fundamentally constrained by the expressiveness of the static, per-frame features they receive as input. Our empirical analyses show that architectural choices in the feature extraction stage, such as the backbone and the fusion mechanism, have a significant impact on overall model performance and are necessary to fully capture the spatial and temporal relationships.

To address this challenge, we propose **ST-Gaze**, a novel architecture designed to generate a more context-aware feature representation for temporal gaze estimation. ST-Gaze is built on three core principles. First, it leverages a convolutional backbone (EfficientNetB3 [35]) to extract rich initial features from separate eye and face patches. The eyes provide the primary signal for gaze direction, while the face offers complementary context, such as head-pose and orientation, which helps disambiguate gaze [5]. Second, we propose a combination of attention mechanisms: a channel-wise attention ECA module [40] adaptively weighs the importance of eye versus face features, while a following self-attention module [39] learns the spatial relationships within the fused feature map. Finally, we introduce a novel recurrence mechanism to address a critical limitation in prior temporal models: *premature spatial pooling*. By collapsing the rich 2D feature map into a single vector before temporal processing, these models irrevocably discard the intra-frame spatial relationships between facial features. Our key innovation is a novel *spatio-temporal recurrence* that models this crucial *intra-frame* context first. We treat the 2D feature map as a spatial sequence and process it with a Gated Recurrent Unit (GRU) [9] to generate a context-aware summary of the frame. This summary, in the form of the GRU's final hidden state, is then propagated through time to model the *inter-frame* dynamics. This design ensures that temporal modelling operates on a spatially informed representation, a capability our ablation study confirms is critical for performance.

We evaluated our proposed ST-Gaze on the EVE dataset [28]. As our work targets video-based gaze estimation for common interactive settings such as desktop and laptop use, EVE is the premier benchmark for this task due to its high-quality, continuous video data of natural user-screen interaction. While other large datasets exist, such as the image-based ETH-XGaze [46] or the fully "in-the-wild" GAZE360 [20], EVE's specific focus aligns directly with our research goals. ST-Gaze establishes a new state-of-the-art for single-view, video-based 3D gaze estimation on EVE with an error of $2.58°$. We limited ourselves to camera-only inputs as, while using screen content can further disambiguate gaze targets [28], we aim to create a more general-purpose 3D gaze estimation model that does not require continuous screen recording. Furthermore, when our model is used as the input to the person-specific adaptation modules of Bao et al. [4], we reduce the final error to $1.87°$ in the offline setting, where all of a person's data is available for calibration, demonstrating the broad utility of our approach.

In summary, our key contributions are:

- We propose ST-Gaze, a novel gaze estimation architecture featuring a spatio-temporal recurrence mechanism that models intra-frame spatial context before propagating information temporally to capture inter-frame context.
- We achieve state-of-the-art performance on the EVE benchmark for video-based 3D gaze estimation, and show that our model serves as a superior backbone for downstream person-specific adaptation methods.
- We provide a comprehensive ablation study that disentangles the contributions of our design choices, including the backbone, attention mechanisms, and the spatio-temporal recurrence, revealing that our proposed recurrence structure and the self-attention module are the most critical drivers of performance.

2. Related Work

Appearance-based Gaze Estimation. This approach aims to directly learn a mapping from images of the face or eyes to a gaze direction, bypassing explicit 3D modelling. Early work demonstrated the feasibility of this approach [43], and the introduction of large-scale datasets like MPIIGaze [45], ETH-XGaze [46] or EVE [28] spurred the development of deep learning solutions. Many methods have explored the optimal source of image information, with some using only eye patches and others arguing for the inclusion of the full face to provide head pose context [44]. A common theme is the fusion of features from multiple streams, such as left and right eye patches, or eyes and the face. However, this fusion is often a simple concatenation followed by a fully-connected layer. More recent works [5, 22, 33] proposed different ways to more intelligently combine eye features, demonstrating that a more sophisticated fusion strategy can yield significant gains. Our work extends this idea by using both channel and self-attention for a more comprehensive fusion of eye and face features

Temporal Models in Gaze Estimation. Gaze is inherently a dynamic process. Recognising this, several works have moved from single-image estimation to using video sequences. The work of Park et al. [28] introduced the EVE dataset and demonstrated that a GRU-based recurrent model (EyeNet-GRU) could leverage temporal consistency to improve upon a static baseline. Similarly, Zhou et al. [47] combined an iTracker-style model with a bidirectional LSTM to capture temporal dynamics. Li et al. [22] proposed a Static Transformer Temporal Differential Network (STTDN) which uses a custom RNN cell to extract sight movement. Ren et al. [33] also incorporated a GRU to learn eye movement dynamics. These works firmly establish the value of temporal information. However, they are all dependent on the quality of the feature vector extracted at each time step. Moreover, they are spatially pooling the features before the temporal module. In contrast to these works which pool spatial features into a single vector, our approach preserves the full spatial feature map. We argue this is critical, as it allows our model to reason about both

intra-frame spatial relationships and inter-frame temporal dynamics, a capability lost in previous methods.

Attention and Transformers for Gaze Estimation. With their success in other vision tasks [31], attention mechanisms and Transformer architectures are emerging in gaze estimation. Cheng and Lu [6] were among the first to explore a hybrid Vision Transformer (ViT) model. Crucially, their work demonstrated that a hybrid approach, which uses a CNN to extract local feature maps before a Transformer models global relationships, outperforms both pure CNN and pure ViT architectures. Nagpure and Okuma [26] used neural architecture search to find an efficient Transformer-based model. More closely related to our work, Ren et al. [33] integrated channel and self-attention modules to enhance feature extraction from fused eye and face features. Our ST-Gaze builds on this but makes a critical architectural distinction. While existing methods typically apply attention and then pool the features to create a single vector for the temporal model, we innovate by preserving the spatial sequence from our self-attention module. By applying a GRU first across this spatial sequence (intra-frame) and then propagating the hidden state across time (inter-frame), we create a recurrent structure that better models the complex spatio-temporal nature of gaze. This approach draws inspiration from architectures in other video understanding tasks, such as ConvGRU [3], and broader tasks tackling spatio-temporal sequences [24]. Our specific approach is, to the best of our knowledge, novel in the context of gaze estimation. Finally, our work also connects to the line of research on person-specific modelling [4, 27]. By providing a much stronger baseline model (ST-Gaze), we show that subsequent person-specific calibration modules like SCPT [4] can achieve even greater performance, highlighting the importance of a powerful, generalized feature extractor.

3. Spatio-Temporal Gaze Network (ST-Gaze) Architecture

Our proposed **Spatio-Temporal Gaze Network (ST-Gaze)** is a video-based gaze estimation network designed to generate a robust spatio-temporal feature representation. For a given frame, the model independently processes the left eye image with the face image, and the right eye image with the face image, yielding two separate gaze predictions. These two predictions for the left and right eye are averaged to produce the final gaze vector. For brevity, the following sections describe the architecture for a single eye-face stream. The overall architecture of ST-Gaze is depicted in Figure 1

Our model consists of four key stages: 1) a multi-stream feature extraction backbone, 2) an attention-based feature fusion module, 3) our novel spatio-temporal recurrence, and 4) a final gaze regression head. We detail each stage below.

3.1. Multi-Stream Feature Extraction

Building on the established success of hybrid CNN-Transformer architectures for gaze estimation [6], the foundation of our model is a dual-stream backbone that processes eye and face patches independently, capturing their distinct characteristics. While many prior works rely on ResNet variants [11, 12, 16, 19, 22, 28, 33], we employ a modified **EfficientNet-B3** [35] as our feature extractor for both streams, selected for its superior performance in other computer vision tasks [2, 35, 37] while still maintaining computational efficiency. Specifically, we adapt the architecture by removing the final average pooling and prediction layers, and we modify the channel dimension in the final blocks to create two separate instances of the encoder:

- **Eye Encoder**: This stream processes a $3 \times 128 \times 128$ eye patch. To maintain a consistent input geometry, images of the right eye are horizontally flipped before being passed to the encoder. The encoder is configured to output a feature map of size $128 \times 8 \times 8$.
- **Face Encoder**: This stream processes a $3 \times 128 \times 128$ face patch. The encoder is configured to output a feature map of size $32 \times 8 \times 8$.

Using separate encoders allows the model to learn specialized features for each modality. For a given frame I_t, the encoders produce an eye feature map $\mathbf{E}_t \in \mathbb{R}^{128\times8\times8}$ and a face feature map $\mathbf{F}_t \in \mathbb{R}^{32\times8\times8}$. Following Ren et al. [33], we choose to represent the eye with more channels ($C_{\text{eye}} = 128$) than the face ($C_{\text{face}} = 32$), based on the established understanding that, while the information from the face helps guide the prediction [5], the eye is still the main contribution to the gaze direction. These maps are then concatenated along the channel dimension to form a unified feature map $\mathbf{X}_t \in \mathbb{R}^{160\times8\times8}$, which serves as the input to the subsequent fusion stage. For our ablation studies, we also implement a variant using ResNet [16] backbones, as detailed in our *ResNetEncoder* implementation.

3.2. Attention-Based Feature Fusion

Simply concatenating features from different domains can be suboptimal, as it may not effectively capture the complex relations between modalities [5]. To fuse the eye and face features, we employ a two-stage attention mechanism as follows

Channel-wise Fusion. First, we apply an Efficient Channel Attention (ECA) module [40] to the concatenated feature map $\mathbf{X}_t$. The ECA module adaptively re-weights the feature channels by learning their interdependencies without dimensionality reduction. This allows the model to dynamically emphasize either eye-specific details or broader facial context based on the input. The kernel size of the 1D convolution within the ECA module is adaptively determined based on the channel dimension $C = 160$, as de-

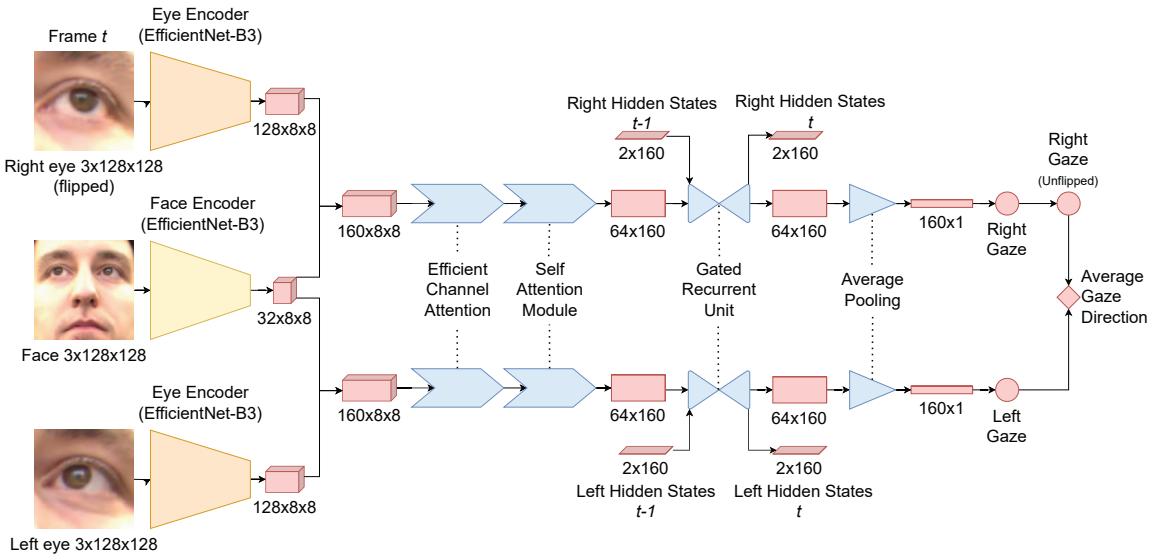


Figure 1. The architecture of our proposed ST-Gaze. An input frame is processed by parallel eye and face encoders. The features are fused via ECA and a Self-Attention Module (SAM). A GRU first processes the feature maps spatially (intra-frame) before propagating its hidden state temporally (inter-frame) to the next frame. A final multi-layer perceptron (MLP) regresses the gaze vector. Input from EVE [28].

scribed in [40]. The output is a channel-recalibrated feature map $\mathbf{X}'_t \in \mathbb{R}^{160\times8\times8}$.

Spatial Feature Learning. Next, we use a **Self-Attention Module (SAM)** [39] to learn long-range spatial dependencies across the feature map. Inspired by the Vision Transformer (ViT) [10] and the work of Cheng and Lu [6] on integrating transformer with gaze estimation, we first reshape $\mathbf{X}'_t$ into a sequence of $S = 64$ patches, where each patch is a feature vector of dimension $D = 160$ and add a learnable positional encoding. This sequence is then processed by a stack of $L = 3$ Transformer encoder blocks with 8 heads each. Each block consists of multi-head self-attention (MHSA) and a feed-forward network. The output of the SAM is a sequence of spatially enriched feature vectors $\mathbf{Y}_t \in \mathbb{R}^{64\times160}$.

3.3. Spatio-Temporal Recurrence

A limitation of prior temporal models for gaze [22, 28, 33, 47] is their reliance on premature spatial pooling. By applying spatial average pooling to the feature map, the entire spatially-aware feature map is collapsed into a single vector, losing the precise arrangement of facial features before the temporal dynamics are considered.

To overcome this, we propose a novel recurrence structure that models *intra-frame* spatial context before modelling *inter-frame* temporal dynamics. We treat the output of our SAM, $\mathbf{Y}_t = \{y_1, y_2, \ldots, y_{64}\}$, as an ordered **spatial sequence**, corresponding to a raster scan of the original 8×8 feature grid. A Gated Recurrent Unit (GRU) [9] with two layers is applied sequentially over these 64 patch vectors. The initial hidden state for this spatial scan at time t, $h_{t,0}$, is the final hidden state from the *previous time step's* spatial scan, $h_{t-1,64}$.

Formally, the process is:

$$h_0^{(0)} = 0^{2\times160} \tag{1}$$

$$\mathbf{Z}^{(t)}, h_{64}^{(t)} = \text{GRU}(\mathbf{Y}^{(t)}, h_0^{(t)}) \quad \text{for } t \geq 0 \tag{2}$$

$$h_0^{(t)} = h_{64}^{(t-1)} \quad \text{for } t \geq 1 \tag{3}$$

Equation 2 describes the intra-frame scan, where the GRU iteratively processes the spatial patches within a single frame to produce a final, context-aware hidden state, $h_{64}^{(t)}$. Equation 3 describes the inter-frame propagation, where this summary state is passed forward in time to initialize the scan for the next frame. This design allows the GRU to first build an understanding of the spatial relationships within a frame before this summary information is propagated through time.

3.4. Gaze Regression and Loss Function

The sequence of spatially-aware features $\mathbf{Z}_t$ from the recurrent module is first aggregated through an average pooling layer across the spatial dimension. This produces a single summary vector $\bar{\mathbf{z}}_t \in \mathbb{R}^{160}$ which is then passed to a two-layer MLP with a Tanh activation function to regress the 3D gaze direction vector $\mathbf{g}_t = (\text{pitch}, \text{yaw})$. The Tanh function constrains the output to $[-1, 1]$, which we then scale by $\pi/2$ to map to the full range of gaze angles in radians. The prediction for the right eye is finally horizontally flipped back.

Loss Function. Our model is trained end-to-end using a weighted combination of losses. The primary loss is the angular error between the predicted gaze vector $\hat{\mathbf{g}}_t$ and the ground-truth vector $\mathbf{g}_t$.

$$\mathcal{L}_{\text{ang}} = \arccos\left(\frac{\hat{\mathbf{g}}_t \cdot \mathbf{g}_t}{\|\hat{\mathbf{g}}_t\|\|\mathbf{g}_t\|}\right) \tag{4}$$

Additionally, following Park et al. [28], we incorporate losses on the predicted Point-of-Gaze (PoG) in both centimetre ($\mathcal{L}_{\text{PoG,cm}}$) and pixel ($\mathcal{L}_{\text{PoG,px}}$) coordinates. The PoG is calculated from the gaze vector using the provided camera geometry, based on the implementation from the EVE framework. The total loss is:

$$\mathcal{L}_{\text{total}} = \lambda_{\text{ang}}\mathcal{L}_{\text{ang}} + \lambda_{\text{px}}\mathcal{L}_{\text{PoG,px}} + \lambda_{\text{cm}}\mathcal{L}_{\text{PoG,cm}} \tag{5}$$

where the λ values are hyperparameters determined experimentally. Finally, we we adopt an offset augmentation strategy during training with a standard deviation of $3°$, following the protocol of Park et al. [28].

4. Experiments

In this section, we present a comprehensive evaluation of our ST-Gaze. We first detail our experimental setup, including the dataset, evaluation metrics, and implementation details. We then compare ST-Gaze against state-of-the-art methods on the EVE benchmark [28]. Finally, we conduct a series of ablation studies to disentangle the contributions of each component of our architecture and validate our design choices.

4.1. Setup

Dataset. We conduct all experiments on the large-scale EVE dataset [28]. We selected EVE as our primary benchmark, as its focus on continuous video of natural user-screen interactions is ideal for evaluating temporal gaze models in desktop and laptop scenarios. The dataset contains over 12 million frames from 54 participants, featuring diverse eye movements and significant head pose variation. Data was collected across four camera views (basler and three webcams at top-left, -centre, and -right) while participants viewed various stimuli (images, videos, Wikipedia pages). We follow the official protocol from Park et al., training on the designated training split. Final performance and key ablations are reported on the test set, with additional detailed results on the validation set available in the supplementary material.

Evaluation Metrics. Following prior work [4, 28], we use the **Angular Error** as our primary evaluation metric, measured in degrees (°) between the predicted 3D gaze vector and the ground-truth vector. As a complementary metric, we also report the Point-of-Gaze (PoG) Error, which is the Euclidean distance in centimetres (cm) between the predicted and actual gaze points on the screen. For fair comparison, all results are reported on the final test set, accessible via the Codalab competition [30] set up by the authors of the EVE dataset.[1] The test set labels are held out, and final performance is evaluated by submitting predictions to the official Codalab competition server.

Implementation Details. Our ST-Gaze is implemented in Pytorch using an EfficientNet-B3 backbone, pre-trained on ImageNet [35]. The model is trained end-to-end for 4 epochs using the Adam optimizer [21] with a batch size of 6 and a base learning rate of $1e^{-4}$, which is scaled linearly with the batch size and follows a cosine annealing schedule. Our total loss is a weighted sum of angular error ($\lambda_{ang} = 1.0$) and PoG error in cm ($\lambda_{cm} = 1e^{-2}$) determined empirically. For experiments with the SCPT modules [4], we use our trained ST-Gaze as a frozen feature extractor and use the publicly available SCPT weights for evaluation.

Model Complexity. We analyse the computational cost and parameter distribution of ST-Gaze to contextualize its performance. For a single-frame, dual-stream forward pass on the input described in Section 3.1, our model computes 6.39 GFLOPs. Additionally, on a PC with a *Nvidia RTX 4090* and an *Intel Core Ultra 9 285K*, it achieves an inference speed of 105 **FPS**, using 800 MB of VRAM. Finally, the vast majority of the **21 M** parameters in our model are concentrated in the dual EfficientNet backbones used for feature extraction (over 94%). In contrast, our proposed combination of attention and spatio-temporal recurrence modules is highly efficient and comprise less than 6% of the total parameters. A detailed breakdown of the parameters' distribution can be found in the supplementary material.

4.2. Comparison with State-of-the-Art

We first evaluate the performance of ST-Gaze as a standalone gaze estimator against several recent and influential methods in video-based gaze estimation. We select baselines that represent different architectural approaches, including the original EVE benchmark model (EyeNet-GRU [28]), Transformer-based (Hybrid-ViT [6]), a similar workflow to ours (FE-NET [33]), and other state-of-the-art architectures (Gaze360 [20]). All baseline results are taken from their original publications or subsequent papers that use the official EVE evaluation protocol, ensuring a fair comparison. As shown in Table 1, ST-Gaze achieves a new state-of-the-art performance on the test set. Our model achieves an angular error of 2.58°, a 25.9% relative improvement over

[1] Competition URL: https://codalab.lisn.upsaclay.fr/competitions/11397

Table 1. Comparison with state-of-the-art methods on the EVE *test* set. All methods listed are appearance-based, using single-view camera input. **Bold** indicates the best performance. *Results reported by [33].

Method	Angular Error (°) ↓	Euclidean PoG Error (cm) ↓
Hybrid-ViT [6]*	3.54	3.92
EyeNet-GRU [28]	3.48	3.85
Gaze360 [20]*	3.45	3.83
FE-NET [33]	3.19	3.55
ST-Gaze (Ours)	**2.58**	**2.87**

the original EyeNet-GRU baseline [28], and a 19.1% relative improvement over FE-NET [33], the best-performing, single-view model on EVE. This result validates that our architecture's focus on superior feature representation translates into substantial performance gains. Moreover, our method achieves comparable performance to other methods that leverage *additional* input modalities. GazeRefineNet [28] and DVGaze [7] both reported an error of 2.49°, but require either the screen content, which may raise privacy concerns for continuous screen capture, or a dual-camera setup, respectively. Our method uses only a single camera view, making it more broadly applicable.

4.3. Performance as a Backbone for Person-Specific Adaptation

Table 2. Performance comparison on the EVE *test* set when used as a backbone for the SCPT adaptation modules [4]. Our model provides a superior initial prediction, leading to a new overall SOTA result.

Backbone + SCPT Modules	Angular Error (°) ↓	
	Offline	Online
EyeNet [28] + SCPT	2.15	1.95
ST-Gaze (Ours) + SCPT	**2.03**	**1.87**
w/o ECA Module	2.15	1.97

To demonstrate the quality of our learned feature representation, we further evaluate its effectiveness as a backbone for person-specific adaptation methods. We use our trained ST-Gaze to provide initial gaze predictions for the Self-Calibration and Person-specific Transform (SCPT) modules from Bao et al. [4]. SCPT is a state-of-the-art unsupervised adaptation technique that refines gaze predictions for a specific person by modelling their unique offsets from a collection of their gaze data, without requiring ground-truth labels. The adaptation can be performed **online**, using only past frames in a sequence, or **offline**, using all available frames for a person to achieve maximum accuracy. As shown in Table 2, our model provides a superior initial prediction, which in turn boosts the performance of SCPT in both online and offline settings We compare against the original EyeNet backbone, which was the baseline used in the SCPT paper, providing a direct and fair comparison. Using ST-Gaze as the feature extractor reduces the error from 2.15° (online) and 1.95° (offline) to 2.03° (online) and 1.87° (offline). These results underscore that a stronger generalized feature extractor is crucial for maximizing the potential of downstream personalization techniques.

4.4. Ablation Study

To validate our architectural design and understand the contribution of each component, we conduct a comprehensive ablation study. We start with our full ST-Gaze and systematically remove or alter key modules. All ablation experiments are evaluated on the EVE validation and test set.

Table 3. Ablation study of the core modules in ST-Gaze on the EVE *test* set. The significant performance drop upon removing the SAM and the GRU, and particularly the degradation from altering the recurrence structure, validates our key design choices. **Bold** indicates the best performance.

Model Configuration	Error ↓	
	Angular (°)	PoG (cm)
ST-Gaze (Full Model)	2.58	2.87
Ablating Fusion Modules:		
w/o ECA Module	**2.56**	**2.86**
w/o SAM	4.84	5.36
Ablating Recurrence:		
w/o GRU	2.88	3.21
Spatial Pooling pre-GRU	2.79	3.13

Impact of Core Architectural Components. Table 3 presents the results of ablating our feature fusion and temporal modelling components. Our full model serves as the baseline, achieving a 2.58° angular error and 2.87 cm PoG error on the test set. The impact of removing the SAM is dramatic, with the angular error (resp. PoG error) increasing to 4.84° (resp. 5.36 cm), highlighting its critical role in learning generalizable spatial relationships. The effect of the ECA module is more nuanced. While its removal results in a marginal 0.02° (resp. 0.01 cm) improvement on the angular error (resp. PoG error), its contribution becomes evident when evaluating the quality of the learned features for downstream adaptation. As shown in Table 2, the features learned *without* ECA are less effective for person-specific adaptation, resulting in a higher final error with the

SCPT modules (1.97° vs. 1.87° for the angular error). This demonstrates that ECA helps produce a more robust feature representation, aligning with our work's primary goal, even if it does not improve raw generalization in this instance.

Disabling the temporal GRU module entirely increases the angular error to 2.88° and PoG error to 3.21 cm, confirming the importance of temporal modelling. Crucially, when we revert to the conventional approach of spatially pooling features *before* the GRU, the angular error increases to 2.79° and the PoG error to 3.10 cm. These results validate our hypothesis: preserving and modelling intra-frame spatial context with our spatio-temporal recurrence leads to lower gaze estimation error than premature spatial pooling.

Table 4. Ablation study on the backbone architecture and fusion strategy on the EVE *test* set. **Bold** indicates the best performance.

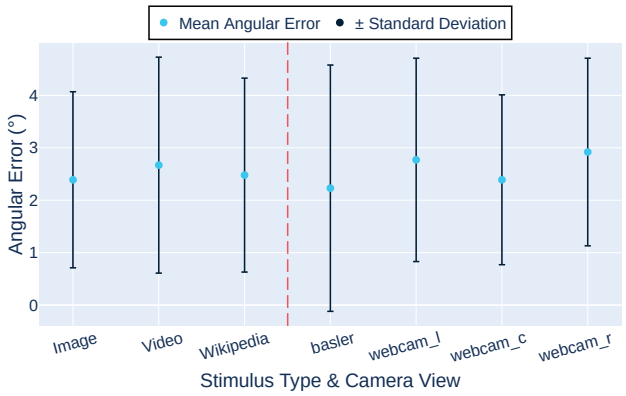
Backbone Configuration	Error ↓	
	Angular (°)	PoG (cm)
EfficientNet-B3 (ST-Gaze)	**2.58**	**2.87**
ResNet-18	4.22	4.99
EfficientNet-B0	2.78	3.09
EfficientNet-B7	2.90	3.24
EfficientNet-V2-S	3.18	3.55
Early Fusion Strategy	3.26	3.66

Impact of Backbone and Fusion Strategy. To validate our specific architectural choices, we conduct further experiments analysing the impact of the backbone and the feature fusion strategy. The results on the test set, presented in Table 4, highlight the importance of these decisions for model generalization. First, we replace the EfficientNet-B3 backbone with several alternatives while keeping the rest of the ST-Gaze architecture constant. With a ResNet-18 [16] backbone, the performance collapses to 4.22°, indicating the failure to generalise. Within the EfficientNet family, our choice of B3 proves to be optimal. A simpler EfficientNet-B0 yields a higher angular error of 2.78°, while a larger and more complex EfficientNet-B7 also sees performance degrade to 2.90°. Similarly, a more recent EfficientNet-V2-S [36] backbone results in a 3.18° angular error. This suggests that simply increasing model capacity does not guarantee better performance, and that EfficientNet-B3 offers the best feature representation for this task, and our fusion and recurrence modules.

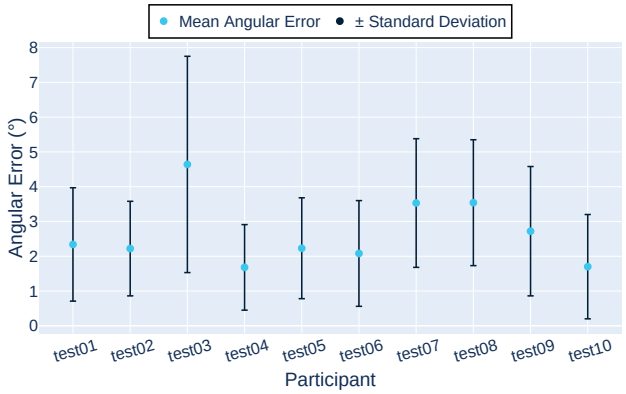
Second, we investigate an early fusion strategy where features from the left eye, right eye, and face are combined before being processed by a single encoder stream. This approach performs poorly, with an angular error of 3.26°. Overall, these ablations confirm that our choice of an EfficientNet-B3 backbone and a late-fusion strategy are critical contributors to the strong generalisation performance of ST-Gaze.

4.5. Qualitative Evaluation

Beyond aggregate metrics, we analyse our model's performance across different conditions and participants to understand its robustness and limitations. Due to the test labels and detailed results being held out, the figures presented for the test performances rely on the mean angular errors and standard deviations provided by the Codalab platform. The detailed breakdown of our results on the validation set can be found in the supplementary material.



(a) Mean angular error and standard deviation across different stimuli (image, video, and wikipedia) and camera views.



(b) Mean angular error and standard deviation across individual participants.

Figure 2. Analysis of ST-Gaze's robustness on the EVE **test set**. (a) Performance is stable across stimuli and camera views, with the centre webcam (*webcam_c*) showing the lowest variance. (b) Performance varies significantly across participants, highlighting the challenge of person-specific factors.

Figure 2 presents this breakdown. Figure 2a analyses performance across different camera views and stimulus types. Performance is largely consistent across the three stimuli, with *Video* content yielding a slightly higher error,

likely due to the more complex and rapid eye movements it elicits compared to static *Image* and *Wikipedia* pages. A more pronounced trend is observed across camera views. While all views perform well, the centrally-mounted *webcam_c* not only achieves low error but also exhibits the lowest variance. This is a promising result for real-world usability, as the *camera_c* position corresponds to where a webcam would typically be placed on a laptop or monitor.

Figure 2b reveals the significant impact of person-specific factors, which remains a central challenge in generalized gaze estimation. The test participants can be broadly clustered into distinct performance groups. A majority of participants (e.g., *test01, test02, test04*) form a high-performance cluster, with mean errors around $2.0°$. Another group (e.g., *test07, test08*) shows moderate performance with errors around $3.0°$. Crucially, one participant, *test03*, is a clear outlier, exhibiting a much higher mean error of $4.6°$. It is precisely for these outlier cases that downstream person-specific adaptation methods like SCPT are essential, as they are designed to learn and compensate for these unique, individual-specific characteristics. Our analysis of the SCPT results confirms this, but with a nuance. For the high-performance cluster, SCPT provides a moderate but consistent error reduction of 20% to 30%. Its impact is most dramatic on the mid-tier performers, such as *test07*, where it reduces the error by 58%. However, for the most challenging outlier, *test03*, the improvement is minimal, at less than 10%. This suggests a key distinction: SCPT is highly effective at correcting for person-specific geometric and anatomical biases, but its effectiveness is limited when the primary source of error stems from persistent issues in the input data. As we will illustrate next, the high error for participant *test03* appears to be caused by such an input challenge.

To visually investigate these performance differences, Figure 3 presents example frames from two representative participants. The four leftmost frames show participant *test03*, the most challenging case in the test set. In each view, a strong blue light reflection on the participant's eyeglasses occludes the eye region. This persistent visual artifact severely degrades the quality of the input features, naturally leading to a high mean prediction error of $4.64°$ and a standard deviation of $3.11°$ as reported on Figure 2b. In contrast, the rightmost frame shows participant *test04*, one of the best-performing participants with a mean prediction error of $1.68°$ and a standard deviation of $1.23°$. The eye region is clearly visible with even lighting, and the pupil's position is unambiguous, allowing our model to extract high-quality features and make an accurate prediction. These examples visually confirm that performance remains heavily dependent on input image quality, with factors like eyeglass reflections being a significant hurdle for all appearance-based methods.

Figure 3. Qualitative examples from the EVE test set [28]. The four leftmost frames show a challenging case (Participant *test03*) while the rightmost frame shows a high-performance case (Participant *test04*). *Images reproduced with permission from authors.*

5. Conclusion

In this paper, we addressed the critical challenge of feature representation in video-based gaze estimation. We proposed the Spatio-Temporal Gaze Network (ST-Gaze), a novel architecture designed to overcome the limitations of premature spatial pooling found in prior temporal models. Our approach integrates the EfficientNet-B3 backbone with a two-stage attention mechanism for effective eye-face feature fusion. A key contribution is our spatio-temporal recurrence, which first models intra-frame spatial context by treating the feature map as a sequence, before propagating this contextual summary through time.

Our experiments on the challenging EVE dataset demonstrate the effectiveness of this design. ST-Gaze achieves a new state-of-the-art angular error of $2.58°$ for single-view 3D gaze estimation. Furthermore, we showed that the superior feature representation learned by our model serves as a powerful backbone for downstream person-specific adaptation, boosting the performance with the SCPT modules to a final error of $1.87°$. Our comprehensive ablation study validated our architectural choices, confirming that the self-attention module and our novel recurrence structure are key drivers of the model's strong generalization performance.

The robust spatio-temporal features learned by ST-Gaze open up several compelling avenues for future research. One promising direction is moving beyond gaze direction to predict higher-level cognitive states, such as cognitive load, confusion, or attention, where the subtle temporal dynamics captured by our model could be beneficial. Furthermore, as our qualitative analysis highlights, a persistent challenge for all appearance-based methods is robustness to visual artifacts like eyeglass reflections. Future work could explicitly tackle this by developing noise-aware training strategies or generative models to reconstruct occluded eye regions, further improving real-world applicability.

6. Acknowledgements

The computing resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government. This research is

supported by Internal Funds KU Leuven (STG/23/054).

References

[1] Andronicus A. Akinyelu and Pieter Blignaut. Convolutional Neural Network-Based Methods for Eye Gaze Estimation: A Survey. *IEEE Access*, 8:142581–142605, 2020. 1

[2] Bhakti Baheti, Shubham Innani, Suhas Gajre, and Sanjay Talbar. Eff-UNet: A Novel Architecture for Semantic Segmentation in Unstructured Environment. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1473–1481, 2020. 3

[3] Nicolas Ballas, Li Yao, Chris Pal, and Aaron Courville. Delving Deeper into Convolutional Networks for Learning Video Representations, 2016. 3

[4] Jun Bao, Buyu Liu, and Jun Yu. The Story in Your Eyes: An Individual-difference-aware Model for Cross-person Gaze Estimation. https://arxiv.org/abs/2106.14183v1, 2021. 1, 2, 3, 5, 6

[5] Yiwei Bao, Yihua Cheng, Yunfei Liu, and Feng Lu. Adaptive Feature Fusion Network for Gaze Tracking in Mobile Tablets, 2021. 2, 3

[6] Yihua Cheng and Feng Lu. Gaze Estimation using Transformer, 2021. 3, 4, 5, 6

[7] Yihua Cheng and Feng Lu. DVGaze: Dual-View Gaze Estimation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 20575–20584, Paris, France, 2023. IEEE. 6

[8] Yihua Cheng, Haofei Wang, Yiwei Bao, and Feng Lu. Appearance-based Gaze Estimation with Deep Learning: A Review and Benchmark. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–20, 2024. 1

[9] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation, 2014. 2, 4

[10] 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, 2021. 4

[11] Arya Farkhondeh, Cristina Palmero, Simone Scardapane, and Sergio Escalera. Towards Self-Supervised Gaze Estimation, 2022. 3

[12] John Gideon, Shan Su, and Simon Stent. Unsupervised Multi-View Gaze Representation Learning. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 4997–5005, New Orleans, LA, USA, 2022. IEEE. 3

[13] E.D. Guestrin and M. Eizenman. General theory of remote gaze estimation using the pupil center and corneal reflections. *IEEE Transactions on Biomedical Engineering*, 53(6): 1124–1133, 2006. 1

[14] D.W. Hansen and Qiang Ji. In the Eye of the Beholder: A Survey of Models for Eyes and Gaze. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(3):478–500, 2010. 1

[15] Katarzyna Harezlak and Pawel Kasprowski. Application of eye tracking in medicine: A survey, research issues and challenges. *Computerized Medical Imaging and Graphics*, 65: 176–190, 2018. 1

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, 2015. 3, 7

[17] Yu He, Shiwei Xie, Han Zhang, and Yang Guanghui Dai. Integrating Eye-Tracking in Games: Enhancing Player Experience and Understanding Social Gaze Issues. In *2024 IEEE 13th Global Conference on Consumer Electronics (GCCE)*, pages 1396–1400, 2024. 1

[18] Shiwei Jin, Ji Dai, and Truong Nguyen. Kappa Angle Regression with Ocular Counter-Rolling Awareness for Gaze Estimation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2659–2668, Vancouver, BC, Canada, 2023. IEEE. 1

[19] Swati Jindal and Roberto Manduchi. Contrastive Representation Learning for Gaze Estimation. https://arxiv.org/abs/2210.13404v1, 2022. 3

[20] Petr Kellnhofer, Adria Recasens, Simon Stent, Wojciech Matusik, and Antonio Torralba. Gaze360: Physically Unconstrained Gaze Estimation in the Wild, 2019. 2, 5, 6

[21] Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, 2017. 5

[22] Yujie Li, Longzhao Huang, Jiahui Chen, Xiwen Wang, and Benying Tan. Appearance-Based Gaze Estimation Method Using Static Transformer Temporal Differential Network. *Mathematics*, 11(3):686, 2023. 2, 3, 4

[23] Erik Lindén, Jonas Sjöstrand, and Alexandre Proutiere. Learning to Personalize in Appearance-Based Gaze Tracking, 2019. 1

[24] Hongbin Liu, Hao Wu, Weiwei Sun, and Ickjai Lee. Spatio-Temporal GRU for Trajectory Classification. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 1228–1233, 2019. 3

[25] Päivi Majaranta and Andreas Bulling. Eye Tracking and Eye-Based Human–Computer Interaction. In *Advances in Physiological Computing*, pages 39–65. Springer, London, 2014. 1

[26] Vikrant Nagpure and Kenji Okuma. Searching Efficient Neural Architecture With Multi-Resolution Fusion Transformer for Appearance-Based Gaze Estimation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 890–899, 2023. 3

[27] Seonwook Park, Shalini De Mello, Pavlo Molchanov, Umar Iqbal, Otmar Hilliges, and Jan Kautz. Few-Shot Adaptive Gaze Estimation. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9367–9376, Seoul, Korea (South), 2019. IEEE. 1, 3

[28] Seonwook Park, Emre Aksan, Xucong Zhang, and Otmar Hilliges. Towards End-to-end Video-based Eye-Tracking. https://arxiv.org/abs/2007.13120v1, 2020. 1, 2, 3, 4, 5, 6, 8

[29] Primesh Pathirana, Shashimal Senarath, Dulani Meedeniya, and Sampath Jayarathna. Eye gaze estimation: A survey on

deep learning-based approaches. *Expert Systems with Applications*, 199:116894, 2022. 1

[30] Adrien Pavao, Isabelle Guyon, Anne-Catherine Letournel, Dinh-Tuan Tran, Xavier Baro, Hugo Jair Escalante, Sergio Escalera, Tyler Thomas, and Zhen Xu. Codalab competitions: An open source platform to organize scientific challenges. *Journal of Machine Learning Research*, 24(198):1–6, 2023. 5

[31] Gracile Astlin Pereira and Muhammad Hussain. A Review of Transformer-Based Models for Computer Vision Tasks: Capturing Global Context and Spatial Relationships, 2024. 3

[32] Christopher Peters, Catherine Pelachaud, Elisabetta Bevacqua, Maurizio Mancini, and Isabella Poggi. A Model of Attention and Interest Using Gaze Behavior. In *Intelligent Virtual Agents*, pages 229–240, Berlin, Heidelberg, 2005. Springer. 1

[33] Guojing Ren, Yang Zhang, and Qingjuan Feng. Gaze Estimation Based on Attention Mechanism Combined With Temporal Network. *IEEE Access*, 11:107150–107159, 2023. 1, 2, 3, 4, 5, 6

[34] Stephen V. Shepherd. Following Gaze: Gaze-Following Behavior as a Window into Social Cognition. *Frontiers in Integrative Neuroscience*, 4, 2010. 1

[35] Mingxing Tan and Quoc V. Le. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks, 2020. 2, 3, 5

[36] Mingxing Tan and Quoc V. Le. EfficientNetV2: Smaller Models and Faster Training, 2021. 7

[37] Mingxing Tan, Ruoming Pang, and Quoc V. Le. EfficientDet: Scalable and Efficient Object Detection, 2020. 3

[38] Yukio Tono. Application Of Eye-Tracking In Efl Learners' Dictionary Look-Up Process Research. *International Journal of Lexicography*, 24(1):124–153, 2011. 1

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need, 2017. 2, 4

[40] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks, 2020. 2, 3, 4

[41] Ahmed Wasfy, Nourhan Anber, and Ayman Atia. Eyeing the Interface: Advancing UI/UX Analytics Through Eye Gaze Technology. In *2024 Intelligent Methods, Systems, and Applications (IMSA)*, pages 538–543, 2024. 1

[42] Tsuyoshi Yamamoto and Kyoko Imai-Matsumura. Teachers' Gaze and Awareness of Students' Behavior: Using An Eye Tracker. *Comprehensive Psychology*, 2:01.IT.2.6, 2013. 1

[43] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. Appearance-based gaze estimation in the wild. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4511–4520, Boston, MA, USA, 2015. IEEE. 1, 2

[44] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. It's Written All Over Your Face: Full-Face Appearance-Based Gaze Estimation. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2299–2308, 2017. 2

[45] Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. MPIIGaze: Real-World Dataset and Deep Appearance-Based Gaze Estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(1):162–175, 2019. 2

[46] 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*, pages 365–381. Springer International Publishing, Cham, 2020. 2

[47] Xiaolong Zhou, Jianing Lin, Jiaqi Jiang, and Shengyong Chen. Learning A 3D Gaze Estimator with Improved Itracker Combined with Bidirectional LSTM. In *2019 IEEE International Conference on Multimedia and Expo (ICME)*, pages 850–855, 2019. 2, 4