Advanced Robotics

ISSN: 0169-1864 (Print) 1568-5535 (Online) Journal homepage: www.tandfonline.com/journals/tadr20

FocusViT: dynamic patch focus for transformer-based gaze estimation

Dan Sochirca, Jouh Yeong Chew & Xucong Zhang

To cite this article: Dan Sochirca, Jouh Yeong Chew & Xucong Zhang (17 Mar 2026): FocusViT: dynamic patch focus for transformer-based gaze estimation, Advanced Robotics, DOI: 10.1080/01691864.2026.2642636

To link to this article: https://doi.org/10.1080/01691864.2026.2642636

Published online: 17 Mar 2026.

ADVANCED ROBOTICS
https://doi.org/10.1080/01691864.2026.2642636

FULL PAPER

OPEN ACCESS Check for updates

FocusViT: dynamic patch focus for transformer-based gaze estimation

Dan Sochirca[a], Jouh Yeong Chew[b] and Xucong Zhang[a]

[a]Intelligent Systems Department, Delft University of Technology, Delft, Netherlands; [b]Research Department, Honda Research Institute Japan, Saitama, Japan

ABSTRACT

Eye gaze information is an important signal for the robot to understand the attention of the human user. Therefore, multiple advanced model architectures have been developed for the gaze estimation task, including the recent vision transformer (ViT). However, due to the patch grid input, vanilla ViTs breaks the fine ocular details into different patches and floods with redundant information from the forehead, cheeks, and background. In this paper, we introduce FocusViT, a lightweight and end-to-end differentiable framework that adapts ViT for the gaze estimation task. It uses a Patch Translation Module to translate patches on informative content dynamically, and then employs a Perturbed Top-K operator to select only the most informative patches for processing. In this way, the proposed method can efficiently use the most informative patches from the full-face image for the gaze estimation task. Our experiments show that combining patch translation and selection reduces the gaze angular error of the ViT model on both the ETH-XGaze and MPIIFaceGaze datasets. Extensive ablation studies confirm that patch translation and token selection are complementary mechanisms that work in synergy to improve model performance.

ARTICLE HISTORY
Received 1 July 2025
Revised 9 November 2025
Accepted 6 February 2026

KEYWORDS
Gaze estimation; transformer; region focus

1. Introduction

Gaze is a unique non-verbal signal in human-robot interaction [1]. It reveals attention, intent, situational awareness, and cognitive load, which are important for the robot to understand human behavior. Therefore, it is critical to detect the human eye gaze for the humanoid robot [2], social robot [3, 4], and human-robot collaboration [5].

Traditional feature-based or model-based gaze estimation usually requires a high-resolution camera and active infrared light sources, which limits its operation to be less than one meter [6]. In contrast, appearance-based gaze estimation, which infers gaze directly from images without requiring explicit eye feature detection, can work with a webcam over a long distance that is suitable for human-robot interaction.

Recent appearance-based methods typically take the entire face as input, requiring the model to reason over both fine-grained ocular cues and global facial context Early convolutional neural network (CNN) models either focused on cropped eye regions or used the entire face as input. Each approach has its trade-offs: using only the eyes captures fine details but ignores global context (e.g. head pose) that might be crucial in low-light or extreme pose settings, whereas using the full face provides context but dilutes the critical eye features. To address this, researchers explored hybrid strategies. For example, multi-region networks process both face and eye patches in parallel to combine local and global cues [7], learnable spatial weights can emphasize eye regions within a full-face CNN [8], and region selection networks dynamically choose the most informative face subregions for each image [9]. These works show that where the model looks in the face is as important as how it learns gaze from the features.

In recent years, *Vision Transformers (ViTs)* [10] have emerged as a powerful alternative to convolutional networks for visual perception tasks. Unlike CNNs, which process images hierarchically through local receptive fields, ViTs treat an image as a sequence of non-overlapping patches (e.g. 16×16 pixels) and embed each patch into a feature vector, referred to as a *token*. A self-attention mechanism is then used to compute pairwise interactions between all tokens, enabling the model to integrate both local and global visual context in a single representation. This global reasoning ability makes ViTs especially suitable for gaze estimation, where subtle eye cues must be interpreted in the context of overall head and facial geometry. However, the patch-based formulation also introduces unique challenges, as we describe next. We identify two key limitations when using vanilla full-face ViTs: (1) patch fragmentation of critical eye

CONTACT Jouh Yeong Chew jouhyeong.chew@jp.honda-ri.com

content, and (2) redundant background and face tokens. We elaborate on the two limitations in the following.

(1) *Patch fragmentation of critical eye content.* The fixed grid patch partitioning of a ViT can split semantically important regions, predominantly the eyes, across multiple patches. If an eye region falls on a patch boundary, its information will be divided into separate tokens, and this can weaken the feature representation and eventually degrade gaze prediction accuracy.
(2) *Redundant background and face tokens.* A full-face image yields many tokens from regions like cheeks, forehead, and background that carry little useful information for the gaze [8, 11]. These redundant tokens not only bring unnecessary computational cost but can also distract the model's attention. Recent efficient transformer studies on image classification tasks show that a substantial subset of tokens can be pruned with negligible impact on final performance []. However, the eye and periocular region have consistently been identified as the primary source of gaze information, while other facial regions have much more subtle or environmentally-subjective impact [8].

To address these shortcomings, we introduce FocusViT, a framework that extends a standard ViT with the dynamic focus of image patches for the gaze estimation task. To combat patch fragmentation, we employ a Patch Translation Module based on a Spatial Transformer Network (STN) [15] that applies a content-aware translation to each patch, allowing them to dynamically recenter on important ocular features. To mitigate token redundancy, we use a differentiable Perturbed Top-K operator [16] that learns to select only the most informative patches for processing. Our entire pipeline is trainable end-to-end. By selecting the image patches instead of feeding all of them into ViT, it enables the model to use smaller patch sizes to improve the performance.

Our experiments demonstrate the effectiveness of FocusViT. On the ETH-XGaze dataset, combining patch translation with Top-K selection reduces the mean angular error (MAE) of a ViT-S baseline from 4.98° to 4.61° – while using only 25% of the original tokens. Our models also show consistent gains on the MPIIFaceGaze dataset, with the best variants reducing the baseline error from 5.72° down to 5.37°.

In summary, our *contributions* are as follows:

(1) We introduce FocusViT, an end-to-end framework that enhances vanilla ViT for gaze estimation by unifying patch translation and selection.
(2) Through extensive evaluation on the ETH-XGaze and MPIIFaceGaze datasets, we show that the proposed method achieves better performance than our baseline methods.
(3) We provide detailed ablation studies showing that patch translation and selection are complementary modules.

2. Related work

2.1. Gaze estimation

Typical appearance-based gaze estimation is mainly dominated by convolutional backbones. iTracker [7] is the first large-scale mobile gaze network, which uses a four-stream CNN (face + both eyes + face grid) trained on the 2.5M frames crowdsourced GazeCapture dataset. [8] employed a Spatial-Weights CNN and showed that using the entire face with learned spatial masks instead of eye-only crops reduced the angular error to 4.8° on EYEDIAP. Later, Dilated-Net [17] preserved high spatial resolution with dilated convolutions and improved accuracy on MPIIGaze, while Gaze360 [18] introduced a panoramic in-the-wild dataset together with a ResNet-GRU temporal model that predicts gaze even when the eyes are off-screen. These CNNs remain strong baselines due to their efficiency and ability to capture local features, but the local receptive fields limit their ability to model long-range dependencies and global context, especially in scenarios involving extreme head poses or dual-camera setups. Transformers and self-attention mechanisms help address these shortcomings by modeling global face context and capturing multi-view or multi-region correspondences. However, most CNN-based or hybrid networks still rely on fixed spatial sampling of features, which does not explicitly resolve the fragmentation of critical eye regions or the redundancy of uninformative facial areas.

The transformer-based breakthroughs started with GazeTR [19], which explored both a pure ViT and a ResNet-ViT hybrid and reported that the hybrid variant already outperformed strong CNN baselines while using fewer parameters. DV-Gaze [20] introduces dual-view Interactive Convolution blocks plus a dual-view Transformer, cutting error by up to 30% under extreme head pose on ETH-XGaze. GazeSymCAT [21] applies cross-attention between left/right eyes and face, and GazeCaps [22] uses self-attention-routed capsules explicitly model inter-eye relations, improving robustness to occlusion and extreme yaw angles. Efficiency-oriented variants like BoT2L-Net [23] insert Bottleneck-Transformer layers into a shallow ResNet and train the network with twin yaw/pitch losses, lowering mean angular error on

Gaze360 without increasing model size. At the opposite end of the spectrum, TransGaze [24] shows that a plain pre-trained ViT can match hybrid models once eye-region tokens are explicitly emphasized, while reducing the training time by half, compared with deep CNNs. Despite these advances, transformer-based gaze estimators typically inherit the fixed patch grid of ViTs and therefore remain susceptible to eye-patch misalignment and the inclusion of redundant background tokens–precisely the two issues we target in this work.

2.2. Patch and token reduction in vision transformers

The computational complexity of Vision Transformers scales quadratically with the number of tokens. To overcome this, an existing area of work tries to improve efficiency by reducing the number of image patches or tokens processed by the model. In ViTs, this can be done pre-transformer or within the transformer. While we found no token reduction work done on gaze estimation, plenty of research has been done on tasks such as image classification and detection.

Pre-transformer selection focuses on shrinking the input sequence externally to the ViT is still scarce. Differentiable Patch Selection [16] learns a perturbed-Top-K mask that crops only the most informative patches and discards the rest, yielding 3-4× FLOP savings while training end-to-end thanks to their differentiable Top-K operator. Our method adopts this operator but uses a smaller patch size and a vanilla ViT as the backbone. STTS [25] uses spatial–temporal scorer ranks tokens across both space and time and keeps only a perturbed-Top-k subset per clip, trimming about 50% of video tokens with negligible loss. AgentViT [26] trains a reinforcement-learning agent to decide on-the-fly which patches to embed. gViT [27] applies Gumbel-Softmax sampling on patches in echocardiography videos. However, pure pre-transformer approaches remain rare, whereas most other works embed the full grid first and prune later.

2.2.1. Intra-layer selection

Zhou et al. [28] differentiates three key token selection mechanisms: token selection based on a scoring function, based on token merging, and based on convolution and pooling. Some prominent examples of scoring-based models are DynamicViT [12], which uses lightweight prediction modules that identify and prune less informative tokens hierarchically across multiple stages. AdaViT [13] with a generalized approach that can skip tokens, heads and blocks in the model with a three-layer decision network. Evo-ViT [29] retains all tokens but uses two computational paths with different computational costs, where unimportant tokens get cheap updates in the model while important tokens get full updates. Instead of dropping tokens, ToMe [14], a token-merging approach, greedily merges the most similar pairs with 2–3× speedups at 0.3% accuracy loss. Pooling approaches such as HVT [30] or PSViT [31] insert pooling or 1-D convolutions between blocks to down-sample tokens in a hierarchical manner.

While these methods effectively improve transformer efficiency, they are designed for generic visual recognition and not for fine-grained tasks such as gaze estimation, where small misalignment of ocular details can drastically affect accuracy. Our approach extends this line of work by coupling token selection with adaptive patch translation, explicitly tailored to preserve eye-centric information.

2.3. Patch translation and deformable sampling

A complementary line of work lets the network move or resize patches instead of (or in addition to) dropping them. Spatial Transformer Network (STN) [15] applies a predicted affine transformation to input feature maps, allowing a network to focus on important regions by translating, scaling, or rotating the patch grid. This approach to differentiable attention to location paved the way for later deformable ViT modules. Deformable Patch-based Transformer (DPT) [32] predicts per-patch offsets + scales based on input content and generates new embeddings using bilinear interpolation from the vanilla patch embeddings. Their method improved ImageNet accuracy and COCO detection with only a slight increase in FLOPs. DAT [33] is a Vision Transformer with Deformable Attention at every layer, that uses learnable offset groups across all queries to shift the keys/values to attend to important regions. In object detection research, Deformable DETR [34] uses a deformable attention module that acts as sparse attention with learned offsets, and leads to faster convergence and reduced complexity from quadratic to linear in the number of patches. A relevant approach to ours is DeBiFormer [35], which combines token selection with spatial adaptation. They introduce Deformable Bi-level Routing Attention, which first finds top-k regions per query (bi-level routing) and then deforms the attended positions within those regions.

In the context of 3D gaze estimation, Zhang et al. [9] introduce a two-stage architecture in which a Region-Selection Network (RSN) first proposes a single, content-dependent crop inside the face image and a subsequent gaze network regresses the gaze vector from that crop. Their method dynamically focuses on visible or well-lit eye regions, outperforms fixed-patch baselines, and proves especially robust under directional illumination,

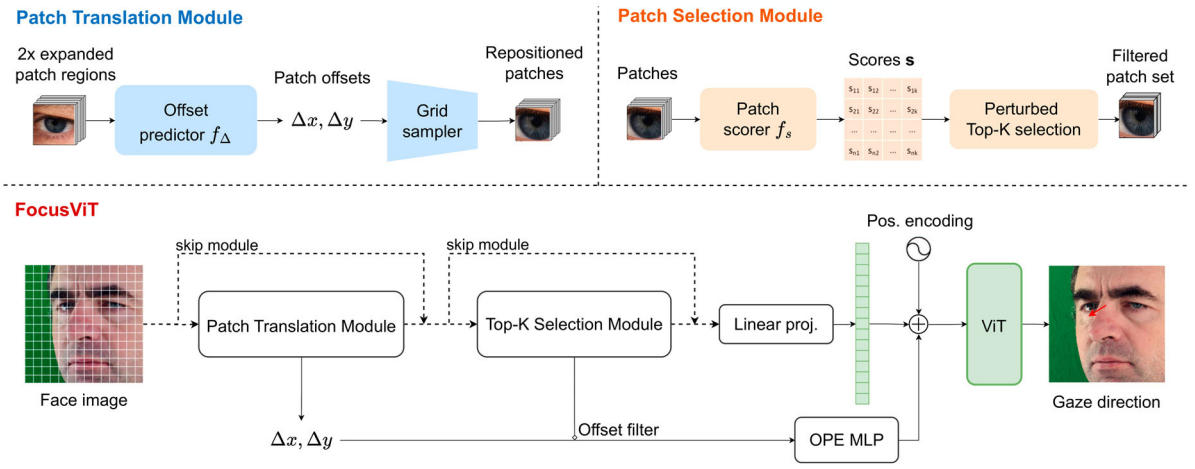

Figure 1. Overview of the proposed FocusViT pipeline. The architecture includes two key modules: the *Patch Translation Module*, implemented using a Spatial Transformer Network (STN) restricted to 2D translation; and the *Patch Selection Module* implemented via a differentiable perturbed Top-*K* operator. The translated and selected patches are then embedded as tokens and processed by a Vision Transformer (ViT) backbone for gaze estimation.

extreme head pose, and partial self-occlusion. Conceptually, this work is close to our work in that it selects face sub-regions. However, it experiments with just up to three selected rectangular crops of 68×68 resolution per image, whereas we keep a set of the most informative 16×16 patches and additionally apply per-patch translations so that each retained token can be centered or move closer to the ocular region. Thus, while prior deformable and region-selection methods demonstrate the benefits of spatial adaptivity, none jointly optimize patch movement and token sparsity within a transformer framework specifically for gaze estimation, which defines the unique contribution of our method.

3. Method

Our FocusViT predicts gaze by coupling two learnable components: a *Patch Translation Module*, implemented as a Spatial Transformer Network (STN) [15], and a *Patch Selection Module*, implemented using a differentiable Perturbed Top-*K* operator [16]. These two modules are placed before a Vision Transformer (ViT) backbone to enable adaptive patch alignment and token selection for efficient gaze estimation (Figure 1).

3.1. Vision transformers

A ViT architecture [10] treats an image as a 1D token sequence that the RGB image is first partitioned into N non-overlapping patches, and each is flattened and linearly projected to a d-dimensional vector. Next, a learnable class token is prepended, positional information is added, and the resulting sequence is processed by L transformer encoder layers with h heads each. Inside the encoders, multi-head self-attention models the pairwise relations among all tokens. For the final prediction of the gaze estimation, the class token is sent to the MLP head to output yaw–pitch angles of the eye gaze, which are then mapped to 3D unit vectors.

3.2. Patch translation module

Our Patch Translation Module consists of Spatial Transformer Network (STN) [15], a learnable component designed to achieve spatial invariance against any spatial transformation. It adaptively transforms the input to a specific pose of interest, with input-dependent parameters of translation, scaling, and rotation, which are part of a 2D affine transformation matrix

$$A = \begin{pmatrix} s_x & r_x & t_x \\ r_y & s_y & t_y \end{pmatrix}, \tag{1}$$

where s, r, and t stand for scale, rotation, and translation, respectively, on horizontal and vertical axes x and y.

The STN consists of three components, including the localization network, the grid generator and the sampler. For an input image I, the localization network predicts the parameters of matrix A. Afterwards, the grid generator uses the predicted parameters to create a mapping from the transformed output image coordinates to the original input image. Finally, the sampler uses the grid generator pixel coordinates and applies bilinear interpolation to extract the output pixel values, which together form the spatially-transformed image.

In our implementation, we restrict the STN to only perform spatial translation. This was done because our datasets are already normalized that images are rotated and cropped using the approach in [36]. The localization

network is a dedicated offset predictor that outputs a 2D translation offset $(\Delta x, \Delta y)$. It is an architecture that uses the first two layers from an ImageNet pretrained ResNet-18 model. Leveraging pretrained weights has improved accuracy in our preliminary experiments.

To predict an offset for each patch, we derived the following process. Although we could simply provide each patch's content of size $3\times16\times16$ to predict the offset, the network is unlikely to accurately infer the context due to the small patch size. Therefore, we provide a larger field of view by extending each 16×16 patch to a 32×32 patch with the same center. This is simply achieved by unfolding the image with a kernel size of 32 and a stride of 16. In this way, the STN can choose a more accurate transformation by seeing 8 extra pixels at each patch border. Due to this design decision, we have to set the scale parameter of the matrix A to 0.5, so that the STN network crops the 32×32 patch in half to the original patch size. The rotation parameters are also set to 0 to disable rotation.

However, there is one issue that remains that applying the patch translation offsets will change the patch content to cause the positional encodings of the ViT to be no longer accurate. To address this, we develop an Offset-aware Positional Embedding (OPE) that works simply by encoding the learned offset to the ViT token dimension. Specifically, For each input patch i with pre-embedding feature $\mathbf{x}_i$, a lightweight MLP predicts a 2-D translation $\mathbf{o}_i = (\Delta x_i, \Delta y_i)$:

$$\mathbf{o}_i = \alpha \tanh\left(W_2\, \sigma\left(W_1\, \mathbf{x}_i\right)\right), \tag{2}$$

where $\sigma(\cdot)$ denotes the GELU activation, *tanh* constrains the offset magnitude, and α scales the maximum shift. Each offset is applied through a restricted Spatial Transformer (translation-only) using bilinear sampling to re-center the patch before linear projection. The OP is trained end-to-end using only the gaze loss, without any explicit offset supervision.

We then and the offset to the standard positional vector to enable the ViT to be aware of the translation

$$ope_i = pe_i + \mathrm{MLP}(\Delta x, \Delta y), \tag{3}$$

where pe_i is the i-th standard positional vector of the ViT model. In our experiments, we try both a one-layer MLP as well as a two-layer variant.

3.3. Patch selection module (Perturbed top-K operator)

We use the ranking-problem formulation to define the patch selection. Specifically, a lightweight ConvNet assigns a relevance score s_i to every patch, and only the K highest-scoring patches are forwarded to the ViT. While it is a simple concept, this hard Top-K operation $arg\_topk(s, K)$ is non-differentiable, since these non-selected patches will always get zero gradient. The standard back-propagation cannot tell the network how to adjust scores. To address this problem, we adopt the perturbed Top-K operator from Cordonnier et al. [16], which is designed to be differentiable based on perturbed optimizers.

The Perturbed Top-K operator works by changing the raw patch relevance scores $s = (s_1, \ldots, s_N)$ with a Monte-Carlo expectation over the hard-Top-K operation after the scores have been noised

$$\begin{aligned} f_\sigma(\mathbf{s}) &= \mathbb{E}_{\boldsymbol{\epsilon}\sim\mathcal{N}(\mathbf{0},\mathbf{I})}\left[\mathbf{TopK}\left(\mathbf{s}+\sigma\boldsymbol{\epsilon},\, K\right)\right] \\ &\approx \frac{1}{T}\sum_{t=1}^{T}\mathbf{TopK}\left(\mathbf{s}+\sigma\boldsymbol{\epsilon}^{(t)},\, K\right), \end{aligned} \tag{4}$$

where $\epsilon^{(t)}$ is i.i.d. Gaussian noise, σ denotes the temperature, and $TopK(\cdot, K)$ returns a 0/1 indicator vector that the K largest components have values $1/K$ to ensure the mask sums to 1. In other words, one-hot Top-K scores are computed after adding noise to the score vector. This operation is repeated T times, and the average is taken. This operation is differentiable because the non-smooth hard-Top-K sits inside the expectation.

In our implementation, we set the parameters similarly to the values in the [16] as $T = 500$ and $\sigma = 0.05$. In our experiments that also use the Patch Translation Module, we set T to 200 due to GPU memory constraints. During training, we follow the same temperature schedule and linearly decay σ to 0. During inference, we set σ to 0, which is equivalent to a deterministic hard Top-K selection. Our patch score network is also based on the implementation in [16] and uses two pretrained ResNet layers as our offset predictor network.

4. Experiment

We first analyze the Patch Translation and Top-K Selection modules in isolation to measure their individual contributions. We then investigate the performance gains from using a small patch resolution. Finally, we evaluate the complete model to demonstrate the complementary benefits of combining both components. Following this analysis, we present the results of our final models on the MPIIFaceGaze dataset and compare them against related work.

4.1. Implementation

4.1.1. Datasets

We train and evaluate our models on the ETH-XGaze [37] and MPIIFaceGaze [8] datasets. ETH-XGaze is a

Table 1. Cumulative ablation study of the patch translation module on ETH-XGaze.

Configuration	MAE (ETH-XGaze)
Baseline	4.98°
+ Offset predictor	**4.71°**
+ Offset-aware Positional Embedding (OPE)	4.72°

high-resolution 1.1 million image dataset of 110 subjects, with extreme head pose and 16 illumination conditions. We do a 50/30 random split of the train partition of 80 subjects to define the train and test sets. MPIIFaceGaze contains 214k images of 15 subjects, collected in the wild with mostly frontal head poses. We use the common cross-subject 15-fold evaluation procedure. For evaluation on ETH-XGaze, we initialize from ImageNet weights. For experiments on MPIIFaceGaze, we initialize model weights by pre-training on ETH-XGaze following previous work [19].

4.1.2. Training settings

All models are implemented by extending the timm-library [38] implementation of Vision Transformers, and trained on a single NVIDIA A40 (48 GB) for 40 epochs. We use the AdamW optimizer and a cosine one-cycle schedule, with the base learning rate of 0.0005. We set the batch size to 200 – if it does not fit the model on the GPU, we set it to the highest number that does.

4.1.3. Evaluation metric

Performance is reported as the mean angular error (MAE) in degrees between predicted and ground-truth gaze vectors.

4.1.4. Baseline

Our reference is the ImageNet pretrained ViT-S model `vit_small_patch16_224` [39] (21M parameters, 16×16 patch size, $N = 196$, $L = 12$, $h = 6$, $d = 384$). The same underlying architecture is used for our models with patch translation and Top-K selection. On ETH-XGaze, the finetuned baseline delivers a mean angular error of 4.98°; on MPIIFaceGaze – 5.72°.

4.2. Contribution of patch translation

In this section, we examine the performance of the patch translation module. We perform a cumulative ablation study on the ETH-XGaze dataset to evaluate different architectural choices within this module. Table 1 summarizes the results.

The key investigated components are:

- *Offset-aware Positional Embedding (OPE)*: Adds the offset-dependent embedding described in Equation (3) to the standard positional encoding.
- *Offset predictor*: A translation module where the offset predictor outputs the offset of each patch. The predictor is initialized with weights pretrained on ETH-XGaze.

The results in Table 1 show a clear progression. Our Default Model immediately improves performance, reducing the MAE from the baseline's 4.98° to 4.71°. However, we can see that adding OPE slightly hurts performance. In this translation-only regime, predicted shifts are small and preserve local topology, so OPE offers no advantage and can slightly regularize features. The role of OPE becomes critical once selection prunes tokens (Section 4.4).

Figure 2 shows a visualization of the best-performing model's behavior (with 4.71° MAE). The figure reveals a consistent pattern that patches located on the forehead and eyebrows shift downwards, while patches near the eye boundary slide to better center the pupil, or the eyelids if not visible. In contrast, there is minimal movement in less informative areas, such as the cheeks and chin, suggesting the module learns to focus on relevant facial features. The visualization also shows that the patch movements, while effective, are relatively modest. This limited motion is likely the reason why the OPE module is not beneficial in this translation-only setup.

4.3. Contribution of learned patch selection

We next evaluate score-based patch selection on a static grid, without patch translation. In this setup, a lightweight scorer network assigns a relevance score to each of the 196 of 16×16 patches, and only the Top-K are passed to the transformer backbone. We test configurations that keep the top 51% (100 patches) and 25% (49 patches) of tokens.

The results, presented in Table 2, show that this approach does not outperform the baseline. The best configuration, which retains 51% of the patches, achieves a MAE of 5.22°, a 0.24° degradation compared to the baseline. This performance decrease is likely because patch fragmentation remains an issue that critical features like the iris can be split across multiple patches. Furthermore, by discarding tokens, the model loses some of the global context it might rely on to compensate for the fragmentation. This helps explain why retaining 51% of patches yields better results than the more aggressive 25% selection.

A closer inspection of the selected patches, shown in Figure 3 (rows 3-4), highlights the learned policy of our

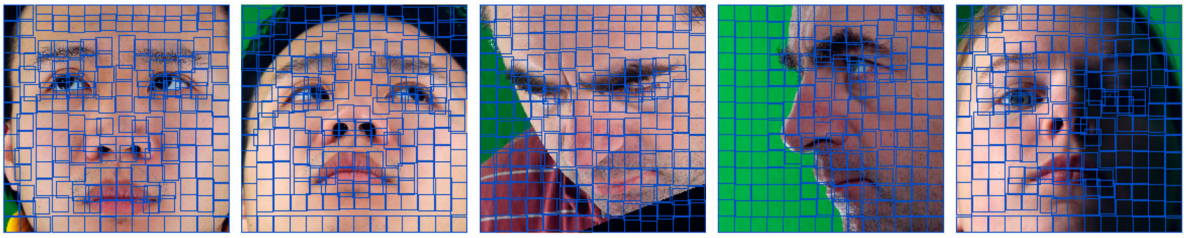

Figure 2. Visualization of the Patch Translation Module on ETH-XGaze. These examples are from our best-performing translation-only model in Table 1 (4.71° MAE). The blue squares show the final patch positions after being dynamically shifted toward more informative regions, such as the eyes and eyebrows, across various head poses.

Table 2. Performance of the learned Top-K selection on a static 16×16 grid (without translation).

Selection %	MAE (ETH-XGaze)
51%	**5.22°**
25%	5.69°

Patch Scorer. The network consistently focuses on the periocular region (eyes, eyelids, eyebrows), while correctly ignoring large, uninformative areas like the cheeks and forehead. However, the scorer network sometimes assigns importance to non-facial artifacts like the subject's hair or background, especially when facial information is sparse.

Furthermore, we observe a bias in the scorer's behavior under uneven lighting. For example, when a face is half-lit, the model tends to select more patches from the darker side. We hypothesize that this is not a preference for darkness itself, but for the high-frequency information and sharp contrast found at the shadow boundaries. These features are rich cues for 3D facial geometry, which the simple convolutional Patch Scorer may have learned to associate with informative regions.

4.4. Effect of joint translation and selection

We now combine our best-performing Patch Translation Module with the Top-K Selection Module to create the complete FocusViT model. This final experiment is designed to show how dynamic patch alignment and learned selection are complementary.

First, we demonstrate the critical importance of Offset-aware Positional Embedding (OPE). As shown in Table 3, activating the Patch Scorer on top of the translated patches without OPE degrades performance, resulting in an error of 6.02°. Adding a hidden layer to the OPE's MLP further refines the performance to 4.61° – a 0.37° improvement over the baseline 4.98°. This confirms that explicitly encoding the patch offsets is essential for the transformer to maintain spatial coherence when patches are both moved and pruned.

Table 3. Performance of the full FocusViT model combining patch translation and selection. The table ablates selection ratio and the effect of Offset-aware Positional Embedding (OPE) in the translation+selection regime.

Configuration	MAE (ETH-XGaze)
25% selection without OPE	6.02°
25% sel. + OPE	**4.61°**
51% sel. + OPE	5.22°

The combined FocusViT model also reveals a major reversal in behavior. In the selection-only experiments (Section 4.3), retaining more tokens (51%) was always better. Here, the opposite is true: the model with more aggressive 25% selection (4.61° MAE) significantly outperforms the 51% version (5.22° MAE). This improvement comes from the effective collaboration between patch translation and token selection. The translation module first 'cleans up' the input by centering patches on informative content like the eyes. As a result, the Patch Scorer is presented with higher-quality candidates and can afford to be more selective, discarding a larger number of tokens without losing critical information, thereby reducing distractions.

A visual analysis in Figure 4 supports this conclusion. Compared to selection on a static grid, the final set of chosen patches is more tightly focused around the periocular region. Patches that could have been selected on the eyebrows are now shifted down to cover the eyes before being selected. This reinforces our core claim: translation and selection are complementary. The translation module creates better candidates, and this allows the selection module to filter more aggressively and effectively.

4.5. Final results and discussion

To conclude our analysis, we summarize the results for our most relevant FocusViT configurations on both the ETH-XGaze and MPIIFaceGaze datasets and compare them against our baseline and related work. The complete results are presented in Table 4.

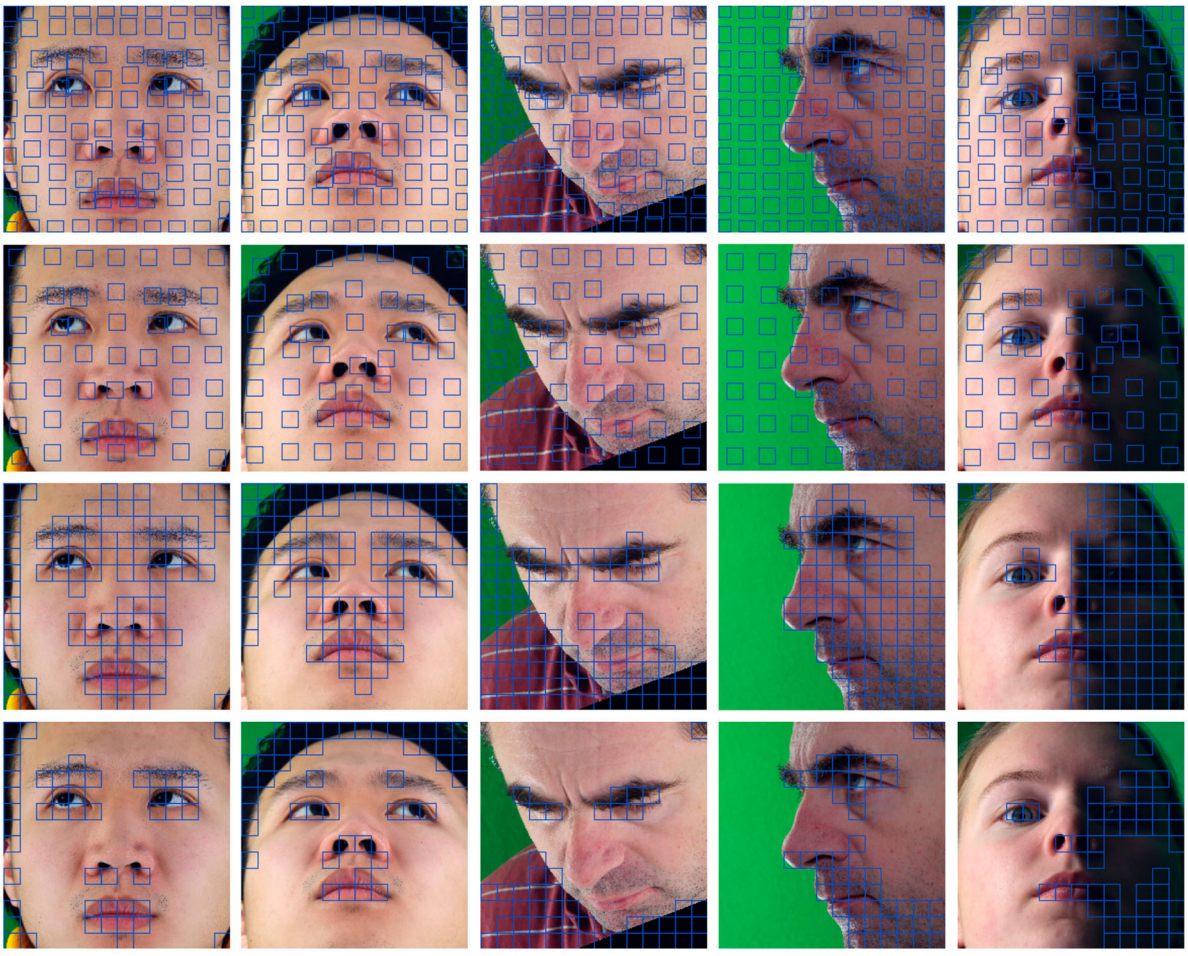

Figure 3. Patch selection by dynamic-grid subsampling (rows 1-2) vs. learned Top-K (rows 3-4). Either 51% or 25% of patches are sampled. The learned Top-K selector focuses on the eyes, nose, mouth, and head contour while consistently avoiding the cheeks and forehead. Some artefacts, such as hair strands or the person's shirt, are sometimes sampled when facial detail is weak. The dynamic-grid subsampling method accurately captures important eye features by shifting the patches, but more irrelevant tokens happen to be selected due to the grid configuration.

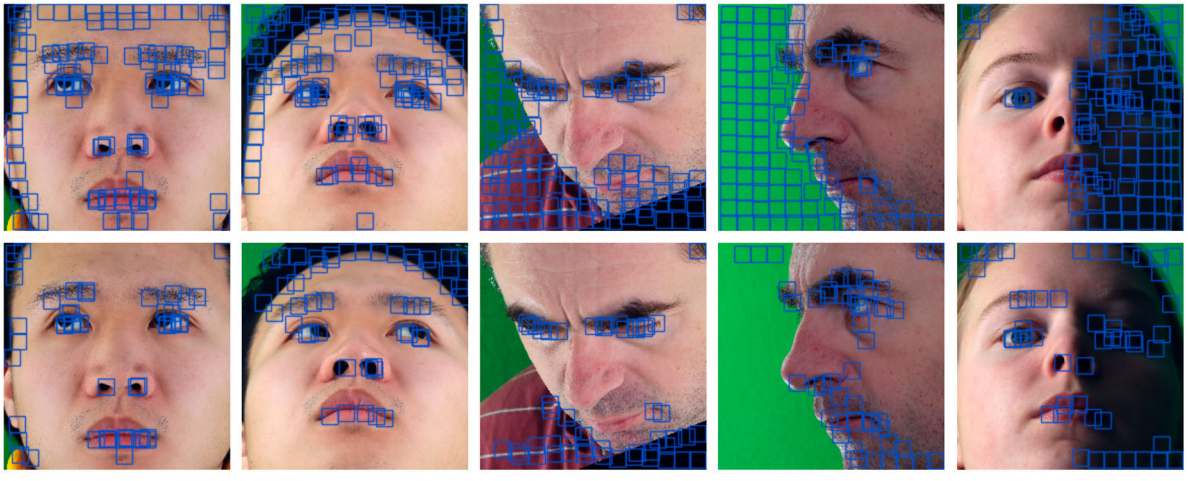

Figure 4. Visualization of the full FocusViT model (25% selection). Patches are first dynamically translated and then scored, leading to a more refined and concentrated selection on the eyes compared to models without translation.

Table 4. Gaze estimation performance on MPIIFaceGaze and ETH-XGaze with different configurations of FocusViT, in comparison with our baseline and transformer-based methods.

Model	Selection %	ETH-X	MPIIFaceGaze
ViT-S (baseline)	–	4.98	5.72
GazeTR [19]	–	4.56	4.96
GazeSymCAT [21]	–	3.28	4.11
UniGaze [40]	–	3.96	4.07
OmniGaze [41]	–	–	2.97
FocusViT	100%	4.72	5.37
FocusViT	51%	5.22	6.25
FocusViT	25%	4.61	5.84

Table 5. Comparison of model scale, parameter efficiency, and performance on ETH-XGaze and MPIIFaceGaze. The table includes representative CNN, hybrid CNN–Transformer, and transformer-based gaze estimators.

Model	Backbone type	#Params (M)
GazeTR [19]	Hybrid (ResNet18 + ViT)	11.4
GazeSymCAT [21]	Hybrid (ResNet50 + ViT)	92.3
UniGaze [40]	Transformer (ViT-H/14)	632.0
OmniGaze [41]	Transformer (ViT-B/16)	86.0
ViT-S (baseline)	Transformer (ViT-S/16)	21.7
FocusViT	Transformer (ViT-T/8)	**5.6**

Our ablation studies on ETH-XGaze revealed clear improvements over the ViT-S baseline (4.98° MAE). Among configurations, combining patch translation and selection significantly improved performance, achieving a 4.61° MAE. The ETH-XGaze dataset is specifically designed to test gaze estimators under large geometric diversity, containing 110 subjects recorded from extreme head poses (±90° yaw/pitch) and 16 illumination settings [37]. On this challenging dataset, our FocusViT achieves 4.61°–4.72° MAE, performing comparably to the hybrid GazeTR model (4.56°) despite using only a lightweight transformer backbone without convolutional pre-processing. This indicates that the proposed patch translation module successfully mitigates the effect of head-pose induced misalignment by dynamically re-centering patches around salient ocular regions, while the selection module suppresses uninformative regions such as cheeks and background. Together, these components enhance geometric robustness without increasing model size or requiring additional supervision.

On the MPIIFaceGaze dataset, our models also showed consistent gains over the 5.72° baseline. The best variants reduced the error significantly. A qualitative analysis of samples from MPIIFaceGaze, shown in Figure 5, confirms that our models behave consistently across datasets. Both the patch translation and selection modules continue to perform well even under challenging low-light conditions. We again note the scorer network's tendency to select patches at high-contrast shadow boundaries, suggesting that it may use these as cues for 3D facial geometry.

4.5.1. Comparison with state-of-the-art methods

We compare our method against recent state-of-the-art gaze estimation models in Table 4, including both transformer-based and hybrid architectures such as GazeTR [19], GazeSymCAT [21], UniGaze [40], and OmniGaze [41]. These methods represent the current frontier of transformer-based gaze estimation, and both UniGaze and OmniGaze employ large-scale pre-training.

While our FocusViT does not surpass the best-performing models such as GazeSymCAT or OmniGaze, it is important to emphasize that our goal differs from these works. Rather than competing on scale or pre-training, FocusViT provides a principled investigation into the structural efficiency of transformers for gaze estimation–specifically, how spatially adaptive patch translation and selection can improve focus and reduce redundancy in transformer representations. This design allows FocusViT to operate with a compact ViT backbone.

In this sense, FocusViT contributes complementary insights to the field, since it highlights how attention-based vision models can be made more efficient and spatially adaptive for fine-grained gaze analysis, forming a foundation that can be integrated into or scaled up within future large transformer-based architectures.

4.5.2. Applicability to hybrid architectures

Although our experiments focus on improving the vanilla Vision Transformer, the proposed Patch Translation and Top-K Selection modules are architecture-agnostic and can be integrated into hybrid CNN–Transformer pipelines. For instance, the translation module could be applied to feature maps from a convolutional encoder before tokenization, while the selection module can be employed to prune tokens in hierarchical transformer stages. Our current goal is to provide a clear understanding of these mechanisms in isolation; extending them to hybrid backbones like GazeTR or Swin-based models is a promising direction for future work.

4.5.3. Parameter efficiency

Table 5 provides a quantitative comparison of model complexity and accuracy across representative gaze estimation methods. Recent hybrid or transformer-based approaches (e.g. GazeTR, GazeSymCAT, UniGaze, and OmniGaze) achieve strong performance but at a substantial computational cost, with model sizes ranging from 11M to over 600M parameters and often relying on large-scale pretraining.

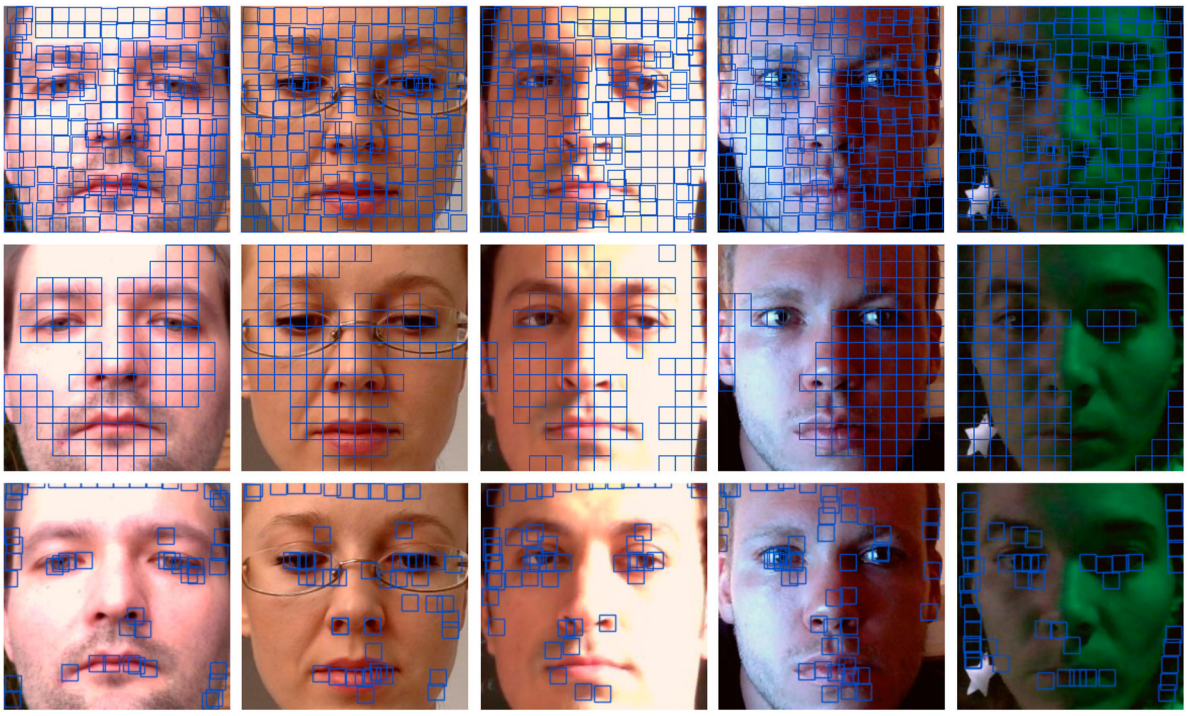

Figure 5. Qualitative results of FocusViT models under various illumination conditions on MPIIFaceGaze. We show visualizations of the translation-only (first row) selection-only models retaining 51% (second row), and the completed FocusViT with 25% of selection patches (third row). Even in challenging low-light conditions, the models consistently attend to the periocular region.

In contrast, FocusViT achieves reasonable performance using only ImageNet pretraining and a compact transformer backbone. For example, FocusViT has merely 5.6M parameters, which is about half of GazeTR parameter count (11.4M). Overall, FocusViT provides a strong trade-off between accuracy and efficiency, demonstrating that lightweight transformer architectures can achieve competitive gaze estimation without resorting to large-scale pretraining or massive model capacities.

4.5.4. Discussion of strengths compared to GazeTR

While FocusViT does not surpass GazeTR in raw estimation accuracy, it offers several complementary advantages. First, it achieves comparable performance using roughly half the parameters (5.6M vs. 11.4M) and without any convolutional backbone. Second, FocusViT provides greater interpretability through patch-level translation and selection maps that reveal where the model attends during gaze prediction (see Figures 2 and 4). Finally, the proposed modules are architecture-agnostic and can be integrated into hybrid CNN–Transformer pipelines to improve efficiency and spatial adaptivity. These aspects highlight the value of 42 our method as a lightweight and transparent alternative to heavier hybrid architectures.

5. Discussion and conclusion

We presented FocusViT, a framework that enhances Vision Transformers for gaze estimation by unifying content-adaptive patch translation with differentiable Top-K selection in a single, end-to-end trainable model. Our experiments successfully validate this approach as a proof of concept, demonstrating that the modules are complementary: translation improves patch quality by centering on ocular features, which in turn allows the selection module to prune distractive tokens more effectively.

For future work, we identify several promising directions:

- *Hybrid Architectures:* Integrating the FocusViT sampling modules into a hybrid CNN-Transformer model could combine the benefits of convolutionally extracted local features with our efficient and targeted attention mechanism.
- *Multi-Scale Patch Selection:* A hierarchical approach with a larger patch size could be explored, where a scorer network evaluates coarse, large-context patches (e.g. 32×32) to select informative regions, while the ViT backbone processes finer-grained patches (e.g. 8×8) from only those areas.
- *Threshold-Based Sparsification:* Replacing the fixed-K selection with an adaptive score threshold could

allow the model to dynamically adjust the number of processed tokens based on scene complexity. This could potentially improve the balance between accuracy and computational cost, and make the selected set of tokens more precise.

Disclosure statement

No potential conflict of interest was reported by the authors.

Notes on contributors

Dan Sochirca was a master student at Delft University of Technology from 2023 to 2025. Prior to it, Dan finished his bechalor study at TU Delft from 2020 to 2023. During his master thesis project, Dan worked on the development of FocusViT as a framework that makes vanilla Vision Transformers more efficient for gaze estimation by dynamically focusing on the most informative facial regions. The result: significantly improved computational efficiency and better gaze prediction accuracy.

Jouh Yeong Chew is a Senior Scientist and Project Leader at Honda Research Institute Japan, where he works on the analysis and modeling of nonverbal cues for cooperative intelligence. He is interested in developing embodied AI to realize a hybrid society in which ubiquitous embodied AI agents and humans coexist in the real world. His research interests include human–robot interaction, machine learning, generative AI, and robotics. He has conducted collaborative research with Toyota Industries Co. and Tadano Ltd. on the development of adaptive support systems for industrial machines, such as forklifts and mobile cranes, by incorporating behavioral intelligence, including gaze saliency. He serves as a Senior/ Guest Editor for Advanced Robotics and has organized workshops at IEEE/ACM conferences such as IROS, ICRA, and HAI. He was also a member of the Advisory Committee of the Innovative Manufacturing, Mechatronics and Materials Forum (iM3F) from 2020 to 2023.

Xucong Zhang is an Assistant Professor at Delft University of Technology. He was a postdoctoral researcher at ETH Zurich from 2018 to 2021, and prior to that completed his PhD at the Max Planck Institute for Informatics in Germany from 2013 to 2018. His research focuses on human-centered artificial intelligence, computer vision, and embodied intelligence, with particular interests in human behavior modeling, gaze estimation, and human–robot interaction. He has published extensively in leading journals and conferences in computer vision and artificial intelligence, and his research has been supported by competitive funding from both public agencies and industry partners.

References

[1] Palinko O, Rea F, Sandini G, et al. Eye gaze tracking for a humanoid robot. In: 2015 IEEE-RAS 15th International Conference on Humanoid Robots (Humanoids). Seoul, South Korea: IEEE; 2015. p. 318–324.

[2] Mutlu B, Kanda T, Forlizzi J, et al. Conversational gaze mechanisms for humanlike robots. ACM Trans Interact Intell Syst. 2012;1(2):1–33. doi: 10.1145/2070719.2070725

[3] Admoni H, Scassellati B. Social eye gaze in human-robot interaction: a review. J Human-Robot Interact. 2017;6(1):25–63. doi: 10.5898/JHRI.6.1.Admoni

[4] Ijuin K, Jokinen KJ, Kato T, et al. Eye-gaze in social robot interactions grounding of information and eye-gaze patterns. In: Proceedings of the Annual Conference of JSAI 33rd (2019). Niigata, Japan: The Japanese Society for Artificial Intelligence; 2019. p. 3J3E402–3J3E402.

[5] Palinko O, Rea F, Sandini G, et al. Robot reading human gaze: why eye tracking is better than head tracking for human-robot collaboration. In: 2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Daejeon, South Korea: IEEE; 2016. p. 5048–5054.

[6] Housholder A, Reaban J, Peregrino A, et al. Evaluating accuracy of the tobii eye tracker 5. In: International Conference on Intelligent Human Computer Interaction. Kent, OH: Springer; 2021. p. 379–390.

[7] Krafka K, Khosla A, Kellnhofer P, et al. Eye tracking for everyone. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition; Las Vegas, NV: 2016. p. 2176–2184.

[8] Zhang X, Sugano Y, Fritz M, et al. 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; Honolulu (HI): 2017. p. 51–60.

[9] Zhang X, Sugano Y, Bulling A, et al. Learning-based region selection for end-to-end gaze estimation. In: 31st British Machine Vision Conference (BMVC 2020). Virtual Event (UK): British Machine Vision Association; 2020. p. 86.

[10] Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: transformers for image recognition at scale; 2020. Preprint. Available from: arXiv:201011929.

[11] Chen H, Liu H, Lan S, et al. Dmagaze: gaze estimation based on feature disentanglement and multi-scale attention; 2025. Preprint. Available from: arXiv:250411160.

[12] Rao Y, Zhao W, Liu B, et al. Dynamicvit: efficient vision transformers with dynamic token sparsification. In: Advances in Neural Information Processing Systems, Virtual: Curran Associates, Inc; Vol. 34; 2021. p. 13937–13949.

[13] Meng L, Li H, Chen BC, et al. Adavit: adaptive vision transformers for efficient image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans (LA): IEEE; 2022. p. 12309–12318.

[14] Bolya D, Fu CY, Dai X, et al. Token merging: your vit but faster; 2022. Preprint. Available from: arXiv:221009461.

[15] Jaderberg M, Simonyan K, Zisserman A, et al. Spatial transformer networks. In: Advances in Neural Information Processing Systems. Montreal (QC): Curran Associates; Vol. 28; 2015.

[16] Cordonnier JB, Mahendran A, Dosovitskiy A, et al. Differentiable patch selection for image recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Virtual Event. IEEE; 2021. p. 2351–2360.

[17] Chen Z, Shi BE. Appearance-based gaze estimation using dilated-convolutions. In: Asian Conference on Computer Vision. Perth: Springer; 2018. p. 309–324.

[18] Kellnhofer P, Recasens A, Stent S, et al. Gaze360: physically unconstrained gaze estimation in the wild. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Seoul: IEEE; 2019. p. 6912–6921.
[19] Cheng Y, Lu F. Gaze estimation using transformer. In: 2022 26th International Conference on Pattern Recognition (ICPR). Montreal (QC): IEEE; 2022. p. 3341–3347.
[20] Cheng Y, Lu F. Dvgaze: dual-view gaze estimation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Paris: IEEE; 2023. p. 20632–20641.
[21] Zhong Y, Lee SH. Gazesymcat: a symmetric cross-attention transformer for robust gaze estimation under extreme head poses and gaze variations. J Comput Des Eng. 2025;12(3):115–129.
[22] Wang H, Oh JO, Chang HJ, et al. Gazecaps: gaze estimation with self-attention-routed capsules. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. Vancouver (BC): IEEE; 2023. p. 2669–2677.
[23] Wang X, Zhou J, Wang L, et al. Bot2l-net: appearance-based gaze estimation using bottleneck transformer block and two identical losses in unconstrained environments. Electronics. 2023;12(7):1704. doi: 10.3390/electronics12071704
[24] Ye L, Wang X, Yao J, et al. Transgaze: exploring plain vision transformers for gaze estimation. Mach Vis Appl. 2024;35(6):128. doi: 10.1007/s00138-024-01609-0
[25] Wang J, Yang X, Li H, et al. Efficient video transformers with spatial-temporal token selection. In: European Conference on Computer Vision. Tel Aviv, Israel: Springer; 2022. p. 69–86.
[26] Cauteruccio F, Marchetti M, Traini D, et al. Adaptive patch selection to improve vision transformers through reinforcement learning. Appl Intell. 2025;55(7):1–26. doi: 10.1007/s10489-025-06516-z
[27] Nilsson A, Azizpour H. Regularizing and interpreting vision transformers by patch selection on echocardiography data. Proc Mach Learn Res. 2024;248:155–168.
[28] Zhou T, Niu Y, Lu H, et al. Vision transformer: to discover the 'four secrets' of image patches. Inf Fusion. 2024;105:102248. doi: 10.1016/j.inffus.2024.102248
[29] Xu Y, Zhang Z, Zhang M, et al. Evo-vit: slow-fast token evolution for dynamic vision transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. Virtual Event. AAAI Press; Vol. 36. 2022. p. 2964–2972.
[30] Pan Z, Zhuang B, Liu J, et al. Scalable vision transformers with hierarchical pooling. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. Virtual Event. IEEE; 2021. p. 377–386.
[31] Tang Y, Han K, Wang Y, et al. Patch slimming for efficient vision transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans (LA); 2022. p. 12165–12174.
[32] Chen Z, Zhu Y, Zhao C, et al. Dpt: deformable patch-based transformer for visual recognition. In: Proceedings of the 29th ACM International Conference on Multimedia. Virtual Event, China: ACM; 2021. p. 2899–2907.
[33] Xia Z, Pan X, Song S, et al. Vision transformer with deformable attention. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. New Orleans (LA): IEEE; 2022. p. 4794–4803.
[34] Zhu X, Su W, Lu L, et al. Deformable detr: deformable transformers for end-to-end object detection; 2020. Preprint. Available from: arXiv:201004159.
[35] BaoLong N, Zhang C, Shi Y, et al. Debiformer: vision transformer with deformable agent bi-level routing attention. In: Proceedings of the Asian Conference on Computer Vision. Hanoi, Vietnam: Springer; 2024. p. 4455–4472.
[36] Zhang X, Sugano Y, Bulling A. Revisiting data normalization for appearance-based gaze estimation. In: Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications. Warsaw: ACM; 2018. p. 1–9.
[37] Zhang X, Park S, Beeler T, et al. Eth-xgaze: a large scale dataset for gaze estimation under extreme head pose and gaze variation. In: Computer Vision–ECCV 2020: 16th European Conference, Proceedings, Part V 16, 2020 Aug 23–28; Glasgow, UK: Springer; 2020. p. 365–381.
[38] Wightman R. Pytorch image models; 2020. [accessed 2025 May 13]. Available at: https://github.com/rwightman/pytorch-image-models.
[39] Wightman R. The timm contributors; 2023. Vision transformer (Small, 16×16) pretrained on ImageNet-1k with AugReg. [accessed 2025 May 13]. Available at: https://huggingface.co/timm/vit_small_patch16_224.augreg_in1k.
[40] Qin J, Zhang X, Sugano Y. Unigaze: towards universal gaze estimation via large-scale pre-training; 2025. Preprint. Available from: arXiv:250202307.
[41] Qu H, Wei J, Shu X, et al. Omnigaze: reward-inspired generalizable gaze estimation in the wild. In: Advances in Neural Information Processing Systems. Virtual: Curran Associates, Inc.; 2025.