

DualStreamFoveaNet: A Dual Stream Fusion Architecture with Anatomical Awareness for Robust Fovea Localization

Sifan Song, Jinfeng Wang, Zilong Wang, Hongxing Wang, Jionglong Su, Xiaowei Ding, Kang Dang

Abstract—Accurate fovea localization is essential for analyzing retinal diseases to prevent irreversible vision loss. While current deep learning-based methods outperform traditional ones, they still face challenges such as the lack of local anatomical landmarks around the fovea, the inability to robustly handle diseased retinal images, and the variations in image conditions. In this paper, we propose a novel transformer-based architecture called DualStream-FoveaNet (DSFN) for multi-cue fusion. This architecture explicitly incorporates long-range connections and global features using retina and vessel distributions for robust fovea localization. We introduce a spatial attention mechanism in the dual-stream encoder to extract and fuse self-learned anatomical information, focusing more on features distributed along blood vessels and significantly reducing computational costs by decreasing token numbers. Our extensive experiments show that the proposed architecture achieves state-of-the-art performance on two public datasets and one large-scale private dataset. Furthermore, we demonstrate that the DSFN is more robust on both normal and diseased retina images and has better generalization capacity in cross-dataset experiments.

Index Terms—Fovea Localization, Vision Transformer, Multi-Cue Fusion

I. INTRODUCTION

THE fovea, a crucial feature of the retina situated at the center of the macula, plays a significant role in producing sharp and detailed vision [1]. Identifying the precise location of the macula and fovea is critical for detecting several retinal diseases, including age-related macular degeneration and diabetic maculopathy [2]–[4]. The distance between the fovea

and abnormalities such as hemorrhages and exudates directly influences the severity of vision loss [4]. Therefore, early detection of the fovea’s location is essential for preventing irreversible damage to vision [3], [5], [6].

Detecting the location of the fovea is crucial in preventing irreversible damage to vision and is an essential part of automated fundus diagnosis [3], [5], [6]. Generally, the fovea is a small, central pit located in the macula region of the retina, responsible for sharp and detailed central vision. In retinal images, the fovea appears as a dark, circular area due to its high density of cone photoreceptors and lack of blood vessels, which enhances visual acuity. It is typically situated approximately 2.5 optic disc diameters temporal (toward the ear) from the center of the optic disc, aligned along the main axis of the retinal blood vessels (Fig. 1-a) [2], [3], [7].

However, several challenges arise in localizing fovea accurately. First, the surrounding retinal tissue’s color intensity makes the fovea’s dark appearance indistinguishable, and there are no anatomical landmarks in its vicinity [4], [7], [8]. Second, retinal diseases such as hemorrhages and microaneurysms can obscure the distinction between the fovea and retinal background (Fig. 1-b and c), leading to inaccurate results [9]–[12]. Similarly, bright lesions like exudates can change the fovea’s lightness to bright and lead to incorrect localization. Third, low light conditions and non-standard fovea locations during photography pose challenges in accurate fovea localization [13]–[15] (Fig. 1-d). Blurred and poorly lit images make estimating macula difficult, and images with centrally located optic discs instead of macula can lead to opposite-side predictions compared to the ground truth. Given these challenges, a robust fovea localization method must take into account global image features to deliver accurate result.

Fortunately, although the image conditions or retina diseases may reduce the distinctness between fovea and macula regions, the global anatomical structures outside the fovea, such as optic disc and blood vessels, have proved to be useful for fovea localization in various studies [2]–[4], [7]–[10], [12], [15]–[20] and other lesion detection research [21], [22]. The fovea center is typically situated along an imaginary line that extends from the center of the optic disc and runs through the macula. This line often aligns with the main superior and inferior temporal blood vessels, which branch out from the optic disc. Experienced clinicians utilize this spatial relationship between the optic disc and main vessels to approximate the fovea position when facing the aforementioned challenging retina

This work was supported by the Key Program Special Fund in XJTU (KSF-A-22) and VoxelCloud, Inc. (Corresponding Authors: Jionglong Su; Xiaowei Ding; Kang Dang)

Sifan Song is with VoxelCloud, Inc., Shanghai, 201114, China (e-mail: ddsong6@gmail.com)

Jinfeng Wang and Jionglong Su are with School of AI and Advanced Computing, XJTU Entrepreneur College (Taicang), Xi’an Jiaotong-Liverpool University, Suzhou, 215123, China (e-mail: Jf.Jacob.Wong@gmail.com, Jionglong.Su@xjtlu.edu.cn)

Zilong Wang is with VoxelCloud, Inc., Shanghai, 201114, China (e-mail: dddwzl3703@163.com)

Hongxing Wang is with School of Big Data and Software Engineering, Chongqing University, Chongqing, 400044, China (e-mail: ihxwang@cqu.edu.cn)

Xiaowei Ding is with Shanghai Jiao Tong University, Shanghai, 200240, China (e-mail: dingxiaowei@sjtu.edu.cn)

Kang Dang is with School of AI and Advanced Computing, XJTU Entrepreneur College (Taicang), Xi’an Jiaotong-Liverpool University, Suzhou, 215123, China and VoxelCloud, Inc., Shanghai, 201114, China (e-mail: Kang.Dang@xjtlu.edu.cn)

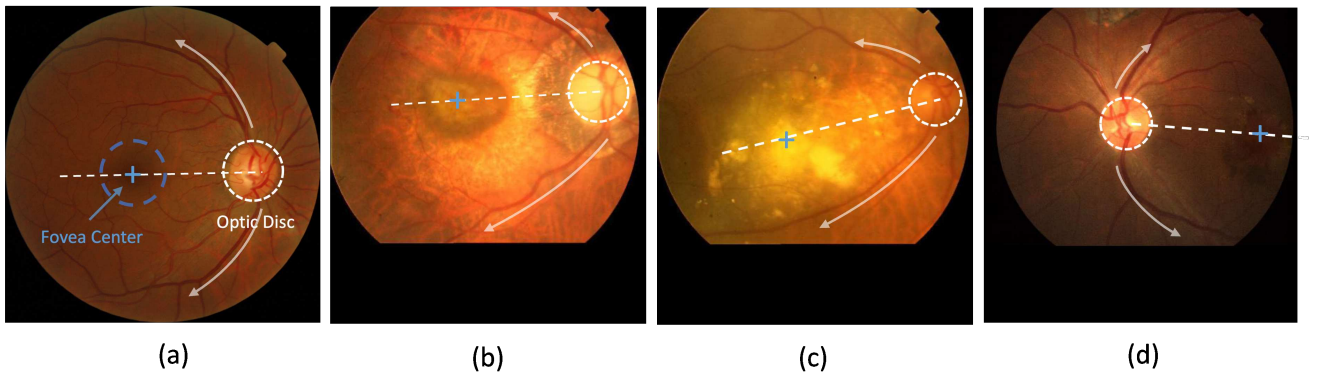


Fig. 1. Illustration of the shared global anatomical relationships among the fovea, optic disc, and main blood vessels, in normal (a), diseased (b, c), and poorly conditioned (d) retina images.

images (Fig. 1-b, c and d) [2], [9], [23].

Previous works have employed carefully designed morphological methods to model the intricate anatomical relationships between the fovea and blood vessels [2]–[4], [9], [12]. However, these traditional methods based on morphological models may falter when faced with images exhibiting rare fovea positions and color intensity variations. While more recent studies have harnessed the power of deep learning to enhance performance, they typically rely solely on fundus images as input [11], [13], [14], [24]–[28] and are implemented on datasets containing few challenging images. These approaches result in three main shortcomings: 1) inadequate exploitation of the anatomical structure outside the macula, as only fundus images are used as input; 2) typical convolution-based architectures lacking the incorporation of global features; and 3) sensitivity to challenging cases, such as rare fovea positions and severe lesions.

To address these challenges, we introduce a novel architecture, DualStreamFoveaNet (DSFN), to adequately utilize the crucial global anatomical relationship (as shown in Fig. 1) for robust fovea localization. For many medical and biomedical research areas, the fusion of multi-branch features has proved to be a crucial technique for integrating information from different cues [29]–[34]. Drawing inspiration from TransFuser [29], our design features a dual-stream encoder that fuses multi-cue features and a decoder that generates result maps. By incorporating images from distinct cues (i.e., fundus and vessel distribution) as inputs for the encoder, we effectively utilize the anatomical structure outside the macula. In the encoder, we merge the multi-cue features of fundus and vessel images using transformer-based modules, dubbed the Bilateral Token Incorporation (BTI), which enables the modeling of global features and long-range connections for fovea localization, ensuring robust performance even in challenging images. Unlike TransFuser, which directly reduces and recovers token numbers by applying average pooling and bilinear interpolation methods—resulting in information loss—we employ TokenLearner and TokenFuser [35] within the BTI module. The attention mechanism of TokenLearner extracts self-learning spatial information from both cues, allowing our design to effectively exploit structural features

along the optic disc and vessel distribution. Furthermore, the attention mechanism reduces the number of tokens in the BTI module, significantly cutting down on computational effort.

Contributions:

- We present a novel approach to fovea localization called DualStreamFoveaNet (DSFN), which utilizes a multi-cue fusion architecture. Unlike traditional convolution-based methods, our approach incorporates transformer-based structures to integrate long-range connections from various cues on a global level.
- We introduce the Bilateral Token Incorporation (BTI) with learnable tokens to enhance the efficiency of transformer-based fusion. This adaptive learning of tokens dramatically reduces the token number from 1024 to 64, while the spatial attention mechanism of the learnable tokens concentrates more on features along the vessel distribution, resulting in robust fovea localization.
- Our proposed DSFN achieves state-of-the-art performance on three datasets (*Messidor*, *PALM*, and *Tisu*) at a mere 25% computational cost (62.11G FLOPs) compared to the best previous work [15] (249.89G FLOPs), demonstrating superior performance and generalization capability in challenging cases.

II. RELATED WORK

A. Anatomical Structure-based Methods

Previous studies have typically relied upon classical image processing methodologies for the estimation of foveal regions. The anatomical correlation between the approximate macular location, optic disc (OD), and blood vessels served as the basis for this reliance [2]–[4], [9], [10], [12], [36]. Specifically, the fovea center is positioned approximately 2.5 OD diameters from the OD’s center, lying on the symmetrical axis of the main vessel branches passing through the OD. These two anatomical landmarks, the distance relationship between the OD and fovea as well as the major vessel orientation, have been widely adopted in previous work for fovea localization relying on image processing techniques [2], [4], [9], [10], [12], [36].

Some studies only rely on the location of OD to identify the foveal region. Narasimha *et al.* [10] propose a two-step

approach that incorporates the distance from the OD center and image intensity to update the region of interest (ROI) before localizing the fovea center. Sekhar *et al.* [8] employ spatial relations to select a sector-shaped candidate ROI. The sector's boundaries extend 30 degrees above and below the line through the image's center and OD. Subsequently, they use a threshold to filter intensity and estimate the foveal region. In color fundus images, blood vessels appear relatively darker in contrast to the OD.

Some studies only utilize the extracted vessels skeleton to estimate the ROI containing the macula. Deka *et al.* [3] and Medhi *et al.* [4] divide the image into several horizontal strips and select the ROI based on the absence of blood vessels in the vicinity of the macula. Thresholding is then utilized to detect the macula. Guo *et al.* [12] propose a morphological technique to fit the segmented skeleton of major vessels using a parabolic curve. The parabola's axis of symmetry is employed to localize the foveal region.

The anatomical relationship between OD and vessels has been utilized in fovea localization. Asim *et al.* [7] estimate the ROI based on the pre-detected OD location and minimum intensity values, excluding the ROI near the vascular tree to enhance accuracy. Li *et al.* [9], Aquino *et al.* [2] and Fu *et al.* [36] also use a parabolic fit of major vessels to detect the orientation of the macula and exploit the anatomical correlation, specifically the distance, between the OD and foveal center for approximate localization. Notably, Fu *et al.* [36] employ a deep learning method (U-Net [37]) rather than image processing techniques to detect the OD and vessels. Nevertheless, the methods only relying on anatomical features may exhibit poor performance when confronted with pathological images. Furthermore, they generally exhibit less competitive performance in comparison to recent deep learning-based approaches.

Efforts have been made to localize the macula without reliance on anatomical features. For example, GeethaRamani and Balasubramanian [11] propose a method to segment the macula using an unsupervised clustering algorithm. Pachade *et al.* [13] directly select the square at the image center as ROI and apply an intensity filter for fovea localization. However, these approaches, which do not consider anatomical features, may fail in cases of varying illumination or when the macula is not situated in the standard location (*i.e.*, the image center).

B. Deep Neural Networks in Fovea Localization

Deep learning has demonstrated superiority over traditional image processing and morphological techniques in many fields of medical image analysis, such as classification, segmentation and object localization [38]–[43]. Regarding the task of fovea localization, existing research can be broadly categorized into two primary domains: regression and segmentation.

Many deep learning-based methods formulate fovea localization as a regression task. Al-Bander *et al.* [25] and Huang *et al.* [27] propose a two-step regression approach, initially predicting the ROI and subsequently inputting them into neural networks for fovea center localization. Meyer *et al.* [26] and Bhatkalkar *et al.* [28] employ pixel-wise distance or heatmap

regression approaches to jointly localize the OD and fovea. Xie *et al.* [14] introduce a hierarchical regression network that integrates a self-attention mechanism in the process of fovea localization [44], [45]. The network predicts the fovea center via a three-stage localization architecture that progressively crops features from coarse to fine.

In addition to regression, deep learning-based methods also leverage the image segmentation paradigm. Tan *et al.* [46] design a single 7-layer convolutional network to point-wise predict the foveal region from input image patches. Sedai *et al.* [24] propose a two-stage image segmentation framework to segment the foveal region progressively. However, standard CNN-based architectures are limited by their fixed-size convolutional kernels, thereby failing to incorporate long-range features adequately. Consequently, these CNN-based architectures may fail under atypical light conditions and non-standard fovea positions, or in the absence of OD and vessel information due to lesions.

To address the constraint of limited receptive field, our prior work proposes a two-branch segmentation architecture (Bilateral-ViT [15]) to model long-range connections. This architecture utilizes a multi-head self-attention (MHSA) mechanism of transformer networks [47]–[49] in a novel way. In Bilateral-ViT, the main branch consisted of 12 consecutive MHSA layers constituting global retinal features, while an auxiliary vessel branch extracted multi-scale spatial information from vessel segmentation. The decoder simultaneously fused multi-cue features from fundus and vessels using multi-scale convolutions. While achieving state-of-the-art results on two publicly available datasets, Messidor [50] and PALM [51], Bilateral-ViT suffers from lacking global multi-cue fusion and high computational costs. To overcome these limitations, we develop a new architecture called DualStream-FoveaNet (DSFN) incorporating a dual-stream encoder capable of globally fusing multi-cue features. Additionally, we introduce adaptively learnable tokens to effectively reduce computation demands.

III. METHODOLOGY

In this work, we introduce DualStreamFoveaNet (DSFN), a novel multi-cue fusion architecture for accurate and robust fovea localization (Fig. 2). DSFN adopts a U-Net-like design with a dual-stream encoder comprising a main stream and satellite stream. Four intermediate Bilateral Token Incorporation (BTI) modules fuse global features from different cues. A decoder with Residual U-blocks (RSU) [41] effectively combines features from both encoder streams. DSFN's transformer-based fusion of multi-cue features via efficient BTI modules maximizes anatomical guidance for accurate and efficient fovea localization.

A. Overall Architecture

The overall architecture of DSFN is depicted in Fig. 2. The encoder consists of two streams: the main stream and the satellite stream, which use ResNet34 and ResNet18 backbones, respectively. The main stream uses the larger ResNet34 to extract rich texture features from fundus images. In contrast,

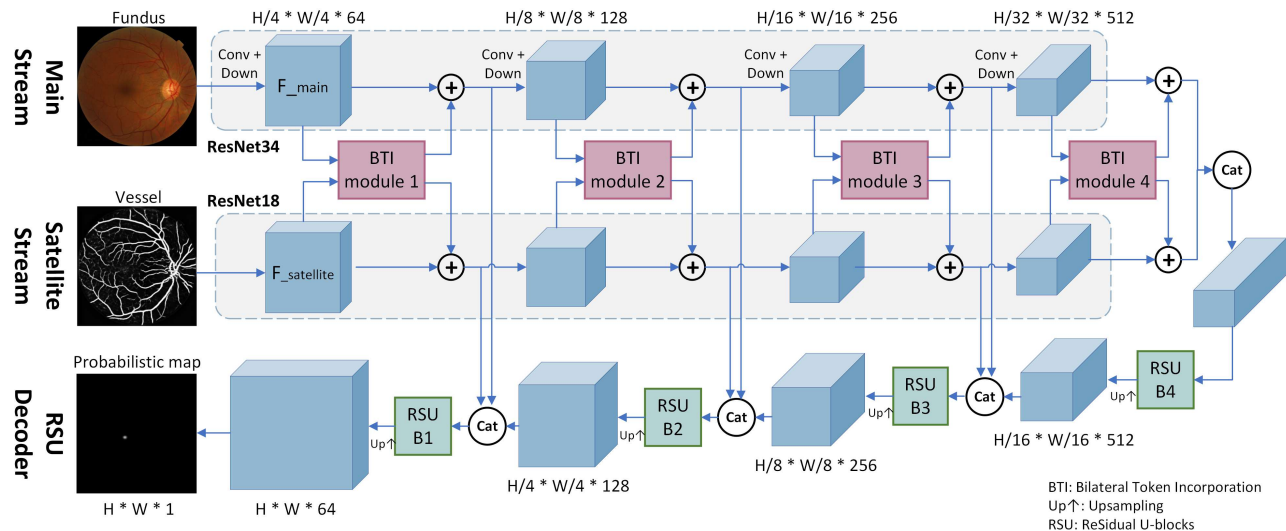


Fig. 2. The overall architecture of our proposed DualStreamFoveaNet (DSFN) network. It consists of a two-stream encoder which effectively incorporates long-range features from both fundus and vessel distributions, and a decoder that generates accurate segmentation results by effective multi-scale feature incorporation.

the satellite stream employs a simpler ResNet18 to extract key and sparse anatomical information from blood vessel distributions. Unlike the main stream that uses fundus images as input, the satellite stream takes a vessel segmentation map generated by a pre-trained model. This pre-trained vessel segmentation model is built on the DRIVE dataset [52] using the TransUNet [49] architecture, which is identical to that used in [15].

The dual-stream encoder, featuring four intermediate modules for multi-cue fusion, is inspired by PVT [53] and TransFuser [29]. As shown in Fig. 2, each stream’s backbone is divided into four convolutional stages, comprising convolution, batch normalization, ReLU, and downsampling layers (Conv+Down). The resulting intermediate tensors (F_{main} and $F_{\text{satellite}}$) are then fed into the BTI module, which includes a modified TokenLearner, T consecutive Multi-Head Self-Attention (MHSA) layers, and a TokenFuser (Fig. 3).

Essentially, we have made four key improvements in the BTI module: 1) We innovatively utilize it to integrate multi-cue features generated from the convolution neural network. After rethinking the spatial mechanism [54], [55] used in TokenLearner, we find it well-suited for anatomical-based feature extraction. We adopt 8×8 learned tokens instead of the original 8 tokens in TokenLearner before each ViT block, as the BTI module needs to adaptively learn features significantly related to anatomical structures (e.g., vessel distribution) for fovea localization. The 8×8 tokens can retain sufficient spatial and other information for subsequent long-range dependency integration. 2) We progressively retain a greater number of channels (128, 256, 512, 1024) for the learned tokens in the deeper stages of the encoder’s BTI module. This approach is based on the observation that as the CNN training progresses, more abstract and high-level information is extracted from each stream. Retaining a greater number of channels allows for better integration and preservation of this information, enabling a gradual and effective fusion process. 3) We con-

catenate the input fundus and vessel features from the two streams first and then use two consecutive 1×1 convolutional layers to remap the features along the channel. Since the non-vessel regions of concatenated vessel features are near zero, they explicitly provide clear clues for the spatial attention map to focus on anatomical structures. 4) After the TokenLearner, we innovatively use MHSA-based feature fusion to address the limitations of CNN-based feature fusion (*i.e.*, limited receptive field and inductive bias of CNN). The MHSA block, through its self-attention mechanism, builds long-range dependencies, facilitating accurate fovea localization using global anatomical landmarks.

In the BTI module, anatomical-related features are adaptively learned, enabling efficient and effective fusion of multi-cue features from both fundus and vessel distributions by the subsequent multi-head self-attention mechanism. In other words, the adaptively learned features reduce the computational power required to encode long-range and global dependencies in the MHSA blocks, making it a practical choice for real-world applications. The output features are element-wise summed with skip-connected features and fed into the next convolutional stage of each stream. Additionally, these output features from the BTI module are forwarded to RSU blocks for decoding operations.

In contrast to the plain convolutional blocks commonly used in basic UNet decoders, the DSFN decoder employs four customized Residual U-blocks (RSU) [41] for feature fusion. Each block is a U-shaped structure for effective multi-scale feature incorporation. Qin et al. [41] demonstrate that the RSU block outperforms other embedded structures (e.g., plain convolution or residual-like, inception-like, and dense-like blocks) due to the enlarged receptive fields of the embedded U-shaped architecture. Furthermore, the superiority of the RSU structure as a decoder for incorporating multiple features has also been assessed by [15]. As illustrated in Fig. 2, RSU B4 serves as the bottleneck block between the encoder and decoder. The

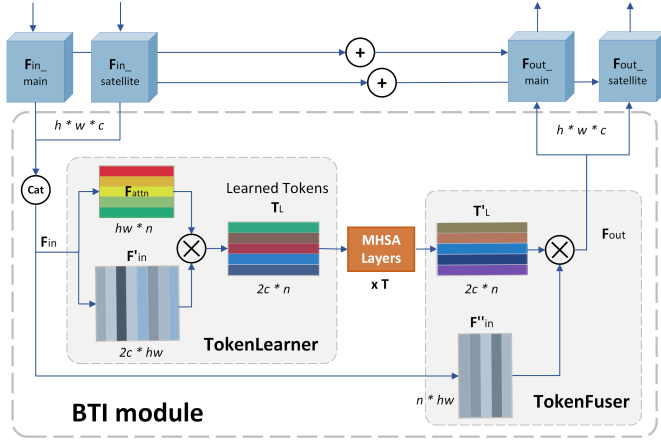


Fig. 3. The structure of BTI module used in our DualStreamFoveaNet (DSFN). It contains TokenLearner, $t \times$ MHA layers and TokenFuser. The h , w and c are height, width, and channel of the corresponding input features. The n represents the number of learned tokens.

input to the other three RSU blocks is a concatenation of three types of features: (i) multi-scale skip-connected features from the main stream, (ii) multi-scale skip-connected features from the satellite stream, and (iii) the hidden feature decoded by the previous RSU block.

B. Bilateral Token Incorporation (BTI) modules

Standard transformer-based architectures, such as Vision Transformer (ViT) [48] and TransUNet [49], typically divide the input image into 2D windows (*e.g.*, a 16×16 grid) to generate tokens. These tokens are then fed into subsequent transformer layers to model long-range feature connectivity. However, the tokens are extracted individually from a fixed-size grid, which may result in uninformative or irrelevant features for visual understanding and increased computational expense. Recent architectures with multiple transformer stages [29], [53] have significantly more layers than the standard ViT, further exacerbating these issues.

To address these challenges, the Bilateral Token Incorporation (BTI) module is introduced in the DSFN architecture (Fig. 3). The BTI module incorporates the TokenLearner [35], which adaptively learns tokens using a spatial attention mechanism. After being processed by Multi-Head Self-Attention (MHSA) layers (with head=8 and layer=12 for each BTI module), the tokens are remapped by the TokenFuser [35] to the original input tensor dimensions (Fig. 3). Consequently, the BTI module offers two main advantages: (1) generating adaptive and learnable tokens to reduce the number of tokens, and (2) effectively merging and fusing features with long-range dependencies from both cues.

1) *TokenLearner*: The TokenLearner modules leverage spatial attention to adaptively generate tokens from the input feature maps for the subsequent MHSA layers. Let $F_{in}^{main} \in \mathbb{R}^{h \times w \times c}$ and $F_{in}^{satellite} \in \mathbb{R}^{h \times w \times c}$ be the input tensors from the two streams, where h , w , and c represent the height, width, and channel of the corresponding BTI module. As shown in Fig. 3, the concatenated feature $F_{in} \in \mathbb{R}^{h \times w \times 2c}$ (Eq. 1) is first fed into the TokenLearner. We customize the

TokenLearner [35] to contain a convolutional block containing two successive point-wise convolutional layers to reduce the dimensionality. This is followed by a flatten and a softmax operation to generate the spatial attention map $F_{attn} \in \mathbb{R}^{h \times w \times n}$ (Eq. 2), where n is the number of learned tokens. Moreover, F_{in} is processed by a point-wise convolutional layer, which is then flattened and transposed to generate a transformed feature map $F'_{in} \in \mathbb{R}^{2c \times h \times w}$ (Eq. 3). Next, the learned tokens generated by spatial attention (F_{attn}) is given in Eq. 4.

$$F_{in} = \text{Concat}(F_{in}^{main}, F_{in}^{satellite}) \quad (1)$$

$$F_{attn} = \text{Softmax}(\text{flatten}(g(F_{in}))) \quad (2)$$

$$F'_{in} = \text{Flatten}(h(F_{in}))^T \quad (3)$$

$$T_L = F'_{in} F_{attn} \quad (4)$$

where $g(\cdot)$ is a block with two point-wise convolutional layers, $h(\cdot)$ represents a point-wise convolutional layer, and the learned tokens are denoted as $T_L \in \mathbb{R}^{2c \times n}$.

Through the spatial attention mechanism, the learned tokens represents an informative combination of features from corresponding spatial locations across the concatenated input cues. Given an input size of 512×512 , in comparison to the 1024 tokens used in ViT and TransUNet [48], [49], we only retain 8×8 tokens for each BTI module. Since the computation of MHSA is quadratic to the number of tokens, the computational cost is significantly decreased. As demonstrated in our experiments, TokenLearner enables us to not only significantly reduce the number of tokens but also extract features related to the distributions of blood vessel for accurate fovea localization.

2) *TokenFuser*: The TokenFuser module reconstructs the tokens processed by the MHSA layers (*i.e.*, T'_L in Eq. 5) back to their original resolution, allowing the output feature (F_{out}) to be further processed by the remaining stages of the DSFN architecture. As shown in Fig. 3, the TokenFuser fuses T'_L with the original input feature map (F_{in}) in the following way:

$$T'_L = \text{MHSA}^{(t)}(\text{MHSA}^{(t-1)}(\dots \text{MHSA}^{(1)}(T_L) \dots)), t = 12 \quad (5)$$

$$F''_{in} = \text{Sigmoid}(r(\text{flatten}(F_{in})^T)) \quad (6)$$

$$F_{out} = \text{Reshape}((T'_L F''_{in})^T) \quad (7)$$

$$F_{out}^{main} = F_{out}[:, :, :c] \quad ; \quad F_{out}^{satellite} = F_{out}[:, :, c:] \quad (8)$$

$$F_{out}^{main} = F_{out}^{main} + F_{in}^{main} \quad ; \quad F_{out}^{satellite} = F_{out}^{satellite} + F_{in}^{satellite} \quad (9)$$

where tensors $F_{in}, F_{out} \in \mathbb{R}^{h \times w \times 2c}$, $F'_{in} \in \mathbb{R}^{n \times h \times w}$ and $T'_L \in \mathbb{R}^{2c \times n}$. $r(\cdot)$ represents a transformation involving two dense layers connected by an intermediate GeLU activation function. As we can see, the intermediate token representation T'_L is remapped to F_{out} which has the same resolution as the original input feature map F_{in} (Eq. 6 and Eq. 7). Next, F_{out} is equally split to F_{out}^{main} and $F_{out}^{satellite}$ (Eq. 8), representing outputs corresponding to the main and satellite streams, respectively. These reconstructed features are then element-wise summed with the skip-connected features from each stream (*i.e.*, F_{in}^{main} and $F_{in}^{satellite}$) to obtain the final output of each stream (F_{out}^{main} and $F_{out}^{satellite}$ in Eq. 9).

IV. EXPERIMENTS

A. Datasets and Network Configurations

We first conduct experiments using the `Messidor` [50] and `PALM` [51] datasets. The `Messidor` dataset was developed for analyzing diabetic retinopathy and comprises 540 normal and 660 diseased retinal images. For this dataset, we utilize 1136 images fovea locations provided by [16]. The `PALM` dataset was released for the Pathologic Myopia Challenge (`PALM`) 2019, which contains 400 images with fovea locations annotated. Of these, 213 images are pathologic myopia images, and the remaining 187 are normal retina images. For fairness of comparison, we follow the same data split as in the existing studies [14] and [15].

Furthermore, we use a large-scale dataset (4103 images, named `Tisu`) which is a subset of data derived from a previously approved study, consists only of de-identified images to ensure subject privacy and data protection. Our current research accesses these images, maintaining confidentiality and ethical compliance. No identifiable subject information is included in the dataset used. Compared to the `Messidor` and `PALM` datasets, `Tisu` is more challenging as it contains a larger number of fundus images with various abnormalities besides hemorrhages, microaneurysms, and exudates. The ground-truth fovea centers are determined by averaging the labels provided by three medical experts. The dataset is split into training and testing sets in a 4:1 ratio.

One of our main contributions is the considerable reduction in FLOPs and GPU usage of our proposed DSFN. Specifically, compared to the previous state-of-the-art Bilateral-ViT [15], our proposed DSFN consumes approximately 25% FLOPs, 48% GPU usage during training and 50% GPU usage during inference. To assess whether the performance of Bilateral-ViT is influenced by its high computational requirements, we also introduce a light version, Bilateral-ViT/Lit (*i.e.*, Bi-ViT/Lit in Table I). The basic architecture of Bilateral-ViT is retained, but we reduce the number of middle channels in each convolutional block by half and decrease the number of tokens from 32×32 to 8×8 . In this case, Bilateral-ViT/Lit demonstrates comparable FLOPs and GPU usage to our DSFN, ensuring a fair evaluation.

B. Implementation Details and Evaluation Metrics

To preprocess the fundus images, we first remove the uninformative black background, and then pad and resize the cropped image region to 512×512 . We simultaneously perform normalization and data augmentation on the input images of the main branch and the vessel branch. To train our DSFN model, we generate circular fovea segmentation masks as the ground-truth from the annotated fovea coordinates. During the inference process, we employ the sigmoid function on the decoder’s output to acquire the segmentation map of the foveal region. The final fovea location coordinates are obtained by calculating the median values of all pixels in the map.

All experiments are implemented in PyTorch and conducted on one NVIDIA GeForce RTX TITAN GPU. We use the Adam optimizer [56] with a batch size of 2. The loss function for predicted segmentation map and ground-truth is a summation

TABLE I
CONFIGURATION COMPARISON OF BILATERAL-ViT MODELS AND THE PROPOSED DSFN.

Methods	Tokens	Image Size	GFLOPs	GPU/T ^a	GPU/I ^b
Bi-ViT [15]	32×32	512^2	249.89	16873	5459
Bi-ViT/Lit [15]	8×8	512^2	83.05	8653	3093
DSFN (Ours)	8×8	512^2	62.11	8083	2727

^a The GPU usage when training (MiB)

^b The GPU usage when inferencing (MiB)

of dice loss and binary cross-entropy. The experimental setup for the DSFN architectures and the splitting of training and test sets on the `Messidor` and `PALM` datasets follow the identical approach as previously reported in [14], [15]. For our proposed DSFN, we set the initial learning rate to $1e^{-3}$, which is gradually decayed to $1e^{-9}$ using the CosineAnnealingLR strategy over 300 epochs on `Messidor` and `PALM`. For the `Tisu` dataset, we set the initial learning rate to $6e^{-5}$ for all related experiments. To reduce overfitting, we adopt a data augmentation strategy to increase the diversity of training data (Details can be found in APPENDIX B). In addition, Xavier weight initialization is utilized for MLP weights in MHSA layers, and normal weight initialization is utilized for MLP biases in MHSA layers.

To evaluate the models’ performance, we employ the R rule as the assessment metric. The R rule is a standard evaluation protocol for fovea localization and widely utilized in many fovea localization studies [14]–[16], [26]. This metric shows the success rate of predictions, where each successful prediction is determined by ensuring that the Euclidean distance between the ground-truth and predicted fovea location coordinates does not exceed a predefined threshold (*i.e.*, optic disc radius R). To provide a comprehensive evaluation, we report the accuracy for different evaluation thresholds from $1/4R$ to $2R$ (*e.g.*, $2R$ indicating that the predefined threshold values are set to twice the radius of the optic disc R). Furthermore, we employ the widely-used mean error (Err) of Euclidean distance between predicted and annotated fovea coordinates to assess the overall performance.

V. RESULTS

A. Comparison to State-of-the-Art

In Table II, we compare the performance of DSFN with existing methods on the public dataset, `Messidor` and `PALM`. Methods are classified based on whether they use deep learning techniques and whether they incorporate multi-cue features. We observe that most traditional morphological methods [2], [16]–[20] rely on landmarks outside the macula, such as vessels or the optic disc. However, in most deep learning-based studies, including coordinate regression methods [14], [25], [27] (\checkmark/C in Table II) and segmentation methods [26], [28] (\checkmark/M), only fundus images are used, resulting in poor incorporation of anatomical relationships throughout the entire image, leading to failure in more challenging cases. Bilateral-ViT [15] is the only previous deep learning-based method that incorporates fundus and vessel features for fovea localization, which achieves better accuracy and robustness.

TABLE II

COMPARISON WITH EXISTING STUDIES USING THE MESSIDOR AND PALM DATASETS BASED ON THE R RULE. THE BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN BOLD AND ITALICS RESPECTIVELY.

Messidor	DL ^a	MF ^b	1/4 R (%)	1/2 R (%)	1R (%)	2R (%)
Gegundez-Arias <i>et al.</i> (2013) [16]		✓	76.32	93.84	98.24	99.30
Giachetti <i>et al.</i> (2013) [17]		✓	-	-	99.10	-
Aquino (2014) [2]		✓	83.01	91.28	98.24	99.56
Dashtbozorg <i>et al.</i> (2016) [18]		✓	66.50	93.75	98.87	99.58
Girard <i>et al.</i> (2016) [19]		✓	-	94.00	98.00	-
Molina-Casado <i>et al.</i> (2017) [20]		✓	-	96.08	98.58	99.50
Al-Bander <i>et al.</i> (2018) [25]	✓ / C		66.80	91.40	96.60	99.50
Meyer <i>et al.</i> (2018) [26]	✓ / M		94.01	97.71	99.74	-
GeethaRamani <i>et al.</i> (2018) [11]			85.00	94.08	99.33	-
Pachade <i>et al.</i> (2019) [13]			-	-	98.66	-
Huang <i>et al.</i> (2020) [27]	✓ / C		70.10	89.20	99.25	-
Xie <i>et al.</i> (2020) [14]	✓ / C		98.15	99.74	99.82	100.00
Bhatkalkar <i>et al.</i> (2021) [28]	✓ / M		95.33	99.74	100.00	-
Bi-ViT (2022) [15]	✓ / M	✓	<i>98.59</i>	100.00	100.00	100.00
Bi-ViT/Lit (2022) [15]	✓ / M	✓	98.50	100.00	100.00	100.00
Wibawa <i>et al.</i> (2023) [23]		✓	-	-	98.24	-
Sigut <i>et al.</i> (2024) [57]		✓	92.69	98.94	99.56	-
DSFN (Ours)	✓ / M	✓	98.86	100.00	100.00	100.00

PALM	DL ^a	MF ^b	1/4 R (%)	1/2 R (%)	1R (%)	2R (%)
Xie <i>et al.</i> (2020) [14]	✓ / C		-	-	94	-
Bi-ViT (2022) [15]	✓ / M	✓	65	83	96	98
Bi-ViT/Lit (2022) [15]	✓ / M	✓	55	80	94	96
DSFN (Ours)	✓ / M	✓	69	85	97	98

^a Whether the method is based on deep learning (DF). The **C** and **M** in this column represent that the method uses **C**oordinate regression and result **M**ap generation, respectively.

^b Whether the method is based on multi-cue features (MF), *e.g.*, fundus images, vessels or optical discs.

TABLE III

COMPARISON OF PERFORMANCE ON NORMAL AND DISEASED RETINAL IMAGES USING THE MESSIDOR AND PALM DATASETS. THE BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN BOLD AND ITALICS RESPECTIVELY.

Messidor	MF ^a	Err _↓	1/4 R(%)		1/2 R(%)		1R(%)		2R(%)	
			Normal	Diseased	Normal	Diseased	Normal	Diseased	Normal	Diseased
UNet (2015) [37]		12.39	95.15	93.33	97.76	95.00	97.95	95.33	97.95	95.33
U2 Net (2020) [41]		7.31	98.51	97.33	<i>99.63</i>	99.50	<i>99.63</i>	99.50	<i>99.63</i>	99.50
TransUNet (2021) [49]		7.61	98.32	97.67	100.00	99.83	100.00	99.83	100.00	99.83
Bi-ViT (2022) [15]	✓	<i>6.81</i>	98.51	98.67	100.00	100.00	100.00	100.00	100.00	100.00
Bi-ViT/Lit (2022) [15]	✓	6.77	<i>98.69</i>	98.33	100.00	100.00	100.00	100.00	100.00	100.00
DSFN (Ours)	✓	6.77	99.07	98.67	100.00	100.00	100.00	100.00	100.00	100.00

PALM	MF ^a	Err _↓	1/4 R(%)		1/2 R(%)		1R(%)		2R(%)	
			Normal	Diseased	Normal	Diseased	Normal	Diseased	Normal	Diseased
UNet (2015) [37]		149.30	74.47	18.87	76.60	41.51	76.60	64.15	78.72	73.58
U2 Net (2020) [41]		62.62	93.62	28.30	<i>95.74</i>	60.38	<i>97.87</i>	84.91	<i>97.87</i>	98.11
TransUNet (2021) [49]		104.38	95.74	18.87	97.87	43.40	<i>97.87</i>	75.47	<i>97.87</i>	84.91
Bi-ViT (2022) [15]	✓	<i>53.70</i>	95.74	37.74	97.87	<i>69.81</i>	100.00	92.45	100.00	96.23
Bi-ViT/Lit (2022) [15]	✓	62.47	87.23	26.42	93.62	67.92	<i>97.87</i>	90.57	<i>97.87</i>	94.34
DSFN (Ours)	✓	48.72	95.74	45.28	97.87	73.58	100.00	94.34	100.00	96.23

^a Whether the method is based on multi-cue features (MF), *e.g.*, fundus images, vessels or optical discs.

The proposed DSFN, which combines a transformer-based multi-cue fusion encoder and adaptive learning tokens, outperforms all previous studies in terms of fovea localization accuracy on the Messidor and PALM datasets. Specifically, in Table II, DSFN achieves the highest accuracy of 98.86% at 1/4R, with gains of 0.71% and 3.53% compared to previous works [14] and [28], respectively. Our network also achieves better performance than Bilateral-ViT [15] and its light version. At evaluation thresholds of 1/2R, 1R, and 2R, DSFN achieves 100% accuracy, indicating a localization error of at

most 1/2R (approximately 19 pixels for an input image size of 512 × 512).

The PALM dataset is more challenging, with a smaller number of images and complex diseased patterns. Our proposed DSFN demonstrates superiority over all other methods on this dataset in Table II, achieving accuracies of 69% and 85% at 1/4R and 1/2R, respectively, which are 4% and 2% better than Bilateral-ViT. In addition, our method achieves a 14% improvement (1/4R) over Bilateral-ViT/Lit, and a 3% improvement (1R) over both Bilateral-ViT/Lit and [14].

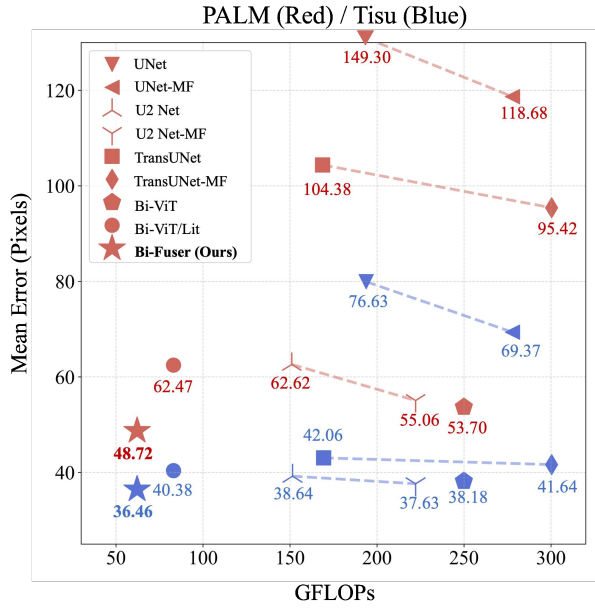


Fig. 4. Visualization of mean errors (Y-axis) of different multi-cue fusion models. X-axis is the computational cost (GFLOPs). The red and blue markers are results on PALM and Tisu datasets, respectively. Numbers below the markers are corresponding mean errors.

Therefore, our DSFN achieves state-of-the-art performance on both Messidor and PALM datasets with high computational efficiency.

B. Fovea Localization on Normal and Diseased Images

In Table III, we separately evaluate the performance of fovea localization for normal and diseased cases in Messidor and PALM datasets to assess the robustness of our method. We compare our proposed DSFN to several widely used segmentation networks, including UNet [37], U2 Net [41], and a hybrid version of TransUNet with ResNet50 for patch embedding [49]. The models of Bilateral-ViT, Bilateral-ViT/Lit and our DSFN are identical to those used in Table II.

In Table III, the proposed DSFN achieves the lowest error (Err) on both datasets. Bilateral-ViT [15] and DSFN both obtain 100% accuracy from $1/2R$ to $2R$ on all the Messidor images, and achieve 100% accuracy of $1R$ and $2R$ on normal PALM images. Compared to existing methods, DSFN demonstrates superior performance on almost all metrics on Messidor. As shown in Table I and Table III, although DSFN has only 25% FLOPs (62.11G) compared to Bilateral-ViT (249.89G), it achieves better performance on diseased images of PALM from $1/4R$ to $2R$, with up to 7.54% improvement compared to the other methods ($1/4R$, Diseased). Its improvement is significantly increased to 18.86% ($1/4R$, Diseased) compared to Bilateral-ViT/Lit (83.05G, *i.e.*, the network with the closest FLOPs to DSFN). Thus, our proposed DSFN is highly reliable in fovea localization on both normal and diseased fundus images with high computational efficacy.

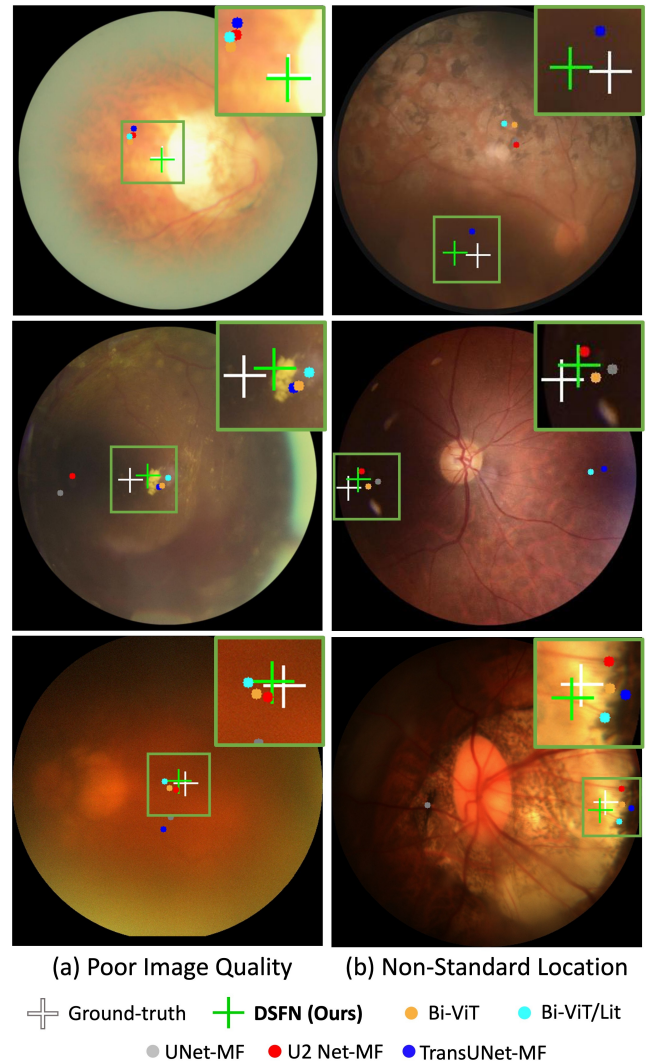


Fig. 5. Comparing visual results of fovea localization predictions by various methods.

C. Comparison of Multi-Cue Fusion Architectures

To comprehensively assess the performance of models with input features from multiple cues (fundus and vessel distributions), we implement a multi-cue fusion version for the baseline models, UNet, U2 Net and TransUNet. We utilize two identical encoders, each with an input of a fundus image and a vessel map. The features are extracted independently and concatenated at the bottleneck for decoding (similar to [29], [58]). These modified baseline models are referred to as UNet-MF, U2 Net-MF and TransUNet-MF. The results in Fig. 4 show that architectures using multi-cue features outperform typical networks with fundus-only input. This is evident in the challenging datasets, PALM and Tisu, where the improvement is pronounced. Moreover, the Tisu dataset is more complex than PALM, with more images (4103 vs. 400) and a wider range of disease types and severity. Therefore, the results on PALM and Tisu demonstrate the potential of architectures that can effectively handle complex datasets.

Fig. 4 shows a comparison of the described architectures (mean error against FLOPs) on PALM (red markers) and

TABLE IV

THE PERFORMANCE OF CROSS-DATASET EXPERIMENTS. **TOP**: THE MODELS TRAINED ON *Tisu* AND TESTED ON *PALM*. **BOTTOM**: THE MODELS TRAINED ON *PALM* AND TESTED ON *Tisu*. THE MODELS USED HERE ARE EXACTLY THE CORRESPONDING ONES USED IN TABLE II AND FIG. 4. THE BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN BOLD AND ITALICS RESPECTIVELY.

<i>Tisu</i> → <i>PALM</i>	Err _↓	1/4 R	1/2 R	1R	2R (%)
UNet-MF	137.14	52.25	64.25	77.25	85.50
U2 Net-MF	69.46	55.00	73.25	90.00	97.25
TransUNet-MF	78.98	53.50	71.75	87.50	95.75
Bi-ViT [15]	64.73	56.00	72.75	90.00	97.75
Bi-ViT/Lit [15]	71.14	55.50	73.25	90.75	97.00
DSFN (Ours)	59.21	55.25	73.25	93.50	98.50
<i>PALM</i> → <i>Tisu</i>	Err _↓	1/4 R	1/2 R	1R	2R (%)
UNet-MF [37]	155.96	19.01	27.74	36.51	44.55
U2 Net-MF [41]	129.86	18.82	30.83	47.96	64.10
TransUNet-MF [49]	118.38	38.26	53.40	68.36	81.04
Bi-ViT [15]	86.92	40.92	54.55	70.97	83.65
Bi-ViT/Lit [15]	91.15	38.02	53.25	70.34	85.40
DSFN (Ours)	74.86	39.34	55.38	75.68	89.42

Tisu (blue markers). Below each marker, we provide the corresponding mean error, and dashed lines connect each standard baseline model with its multi-cue fusion architecture (MF). The multi-cue fusion versions (UNet-MF, U2 Net-MF, and TransUNet-MF) outperform their standard versions at a considerably higher computational cost due to the additional encoder. For all multi-cue fusion architectures, the proposed DSFN achieves the best results with the smallest errors on both *PALM* (48.72 pixels) and *Tisu* (36.46 pixels). Furthermore, it demands only 62.11G FLOPs, which is four times less than the FLOPs required by the next best-performing approach, Bilateral-ViT (249.89G). Compared to the model with comparable FLOPs (Bilateral-ViT/Lit, 83.05G), DSFN shows significant advantages of 13.75 and 3.92 on *PALM* and *Tisu*, respectively.

Fig. 5 provides visual results of fovea localization on images with severe diseases from the *PALM* and *Tisu* datasets. These images in Fig. 5-a and Fig. 5-b suffer from poor image quality and non-standard fovea locations, respectively. Our DSFN generates the most accurate predictions for several challenging cases with poor lighting conditions and blurred appearance (Fig. 5-a). For another challenging types (Fig. 5-b), where the macula is close to the image boundary, the predictions of DSFN (green crosses) are closest to the ground-truth (white crosses). On the contrary, predictions from alternative architectures lacking the ability to globally incorporate long-range multi-cue features may lead to fovea locations appearing on the incorrect side of the optic disc (Fig. 5-b). These results suggest that DSFN can adequately model fundus and vessel features of two streams, resulting in superior performance compared to other networks.

D. Performance of Cross-Dataset Experiments

In our cross-dataset experiments evaluating the generalization capabilities of the proposed DSFN, we focus on two demanding datasets—*PALM* and *Tisu*, characterized by a higher prevalence of retinal diseases and lower image quality. These datasets were interchangeably assigned as training and

TABLE V

COMPARISON OF PERFORMANCE BETWEEN DIFFERENT INPUTS FOR THE MAIN AND SATELLITE STREAMS ON *PALM* AND *Tisu*. THE BEST AND SECOND BEST RESULTS ARE HIGHLIGHTED IN BOLD AND ITALICS.

<i>PALM</i>	Err _↓	1/4 R	1/2 R	1R	2R (%)
Fundus+Vessel (Ours)	48.72	69	85	97	98
Fundus-only	54.46	64	85	95	98
Vessel-only	72.25	57	75	92	97
Fundus+Vessel*	49.61	67	84	95	98
<i>Tisu</i>	Err _↓	1/4 R	1/2 R	1R	2R (%)
Fundus+Vessel (Ours)	36.46	52.32	75.49	92.93	97.44
Fundus-only	37.48	52.32	73.78	91.10	96.95
Vessel-only	48.13	33.66	61.83	89.39	96.83
Fundus+Vessel*	37.34	51.59	73.66	91.95	97.44

* denotes that the input of the satellite stream is pre-trained and generated by UNet (*i.e.*, a weaker cue).

test sets. Notably, in Table IV-**Top**, DSFN exhibits substantial improvements, achieving reductions of 5.52 and 11.93 pixels in average localization error at the original image resolution compared to the previously top-performing method, Bilateral-ViT, and its lighter version, respectively. Additionally, in Table IV-**Bottom**, where models were trained on a smaller dataset (*PALM*) and tested on a larger, more challenging one (*Tisu*), DSFN demonstrates significant enhancements of 12.06 and 16.29 pixels in average location error compared to Bilateral-ViT and its lighter version, respectively. These findings underscore the DSFN’s adaptability to diverse and previously unseen cases, mitigating overfitting concerns and highlighting its robust generalization capability.

E. Ablation Study

1) *Comparison of Inputs for DSFN*: Table V compares the performance of DSFN when using different inputs. The standard input configuration uses fundus images and vessel maps as inputs for the main and satellite streams, respectively (Fundus+Vessel). These experiments achieve the best accuracy on all metrics, with the smallest mean error on both *PALM* (48.72 pixels) and *Tisu* (36.46 pixels). When using fundus images as the second input (Fundus+Fundus), the model’s performance slightly degrades as feeding fundus images into the satellite stream does not provide the explicit anatomical structure for the BTI module to learn where to focus its attention. Experiments using vessel maps as both inputs (Vessel+Vessel) lead to significant accuracy decreases on both *PALM* and *Tisu* by 23.53 and 11.67 pixels, respectively, indicating a severe loss of information. In addition, we conduct experiments where the input of the satellite stream (*i.e.*, vessel distribution map) is pre-trained and generated using a weaker architecture, UNet (Fundus+Vessel*), as opposed to TransUNet in the default setting (Fundus+Vessel). These experiments on *PALM* and *Tisu* demonstrate the performance of our DSFN when the vessel input serves as a weak cue. The outcomes presented in Table V suggest that the performance of Fundus+Vessel* generally falls between that of Fundus+Vessel and Fundus-only. This finding demonstrates that even when providing a weak cue, the DSFN effectively learns valuable anatomical-related features, instead of starting to learn spatial attention from scratch (*i.e.*, Fundus-only).

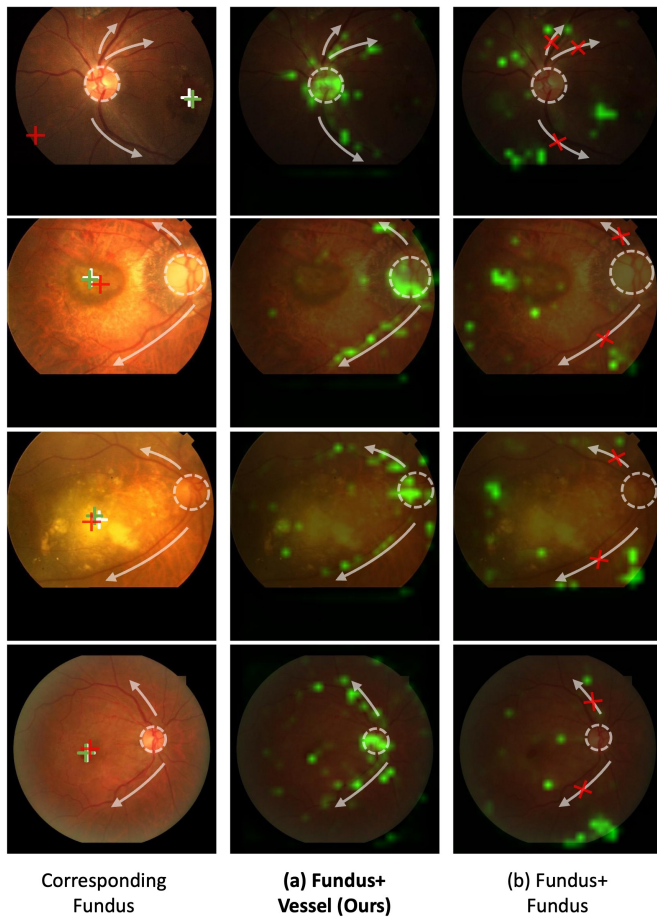


Fig. 6. Visualizations generated from spatial attention maps, indicating the focus of TokenLearner in BTI module. These visual results have been resized and superimposed onto the corresponding fundus. Circles with dotted lines represent the position of optic discs, and arrows represents the direction of main vessel branches. The white, green, and red crosses in the first column are the ground-truth, prediction result with the second input of vessel (a) and fundus (b), respectively.

We utilize spatial attention maps to elucidate the self-learning features extracted by TokenLearners, visually demonstrating these weight maps in Fig. 6. To visualize the attention of tokens in an element-wise manner, we maximize the saliency values among all the tokens n for each position $h \times w$ of the spatial attention map $F_{\text{attn}} \in \mathbb{R}^{h \times w \times n}$ (refer to Eq. 2). Subsequently, we normalize these values and superimpose the resulting maps onto the corresponding fundus, facilitating a comparison of their structural relationships.

In our experiments using Fundus+Vessel inputs, the spatial attention maps focus on structural features along the optic disc and the direction of vessel branches (Fig. 6-a). This focus is feasible and consistent to the significant anatomical relationships between these structures and the fovea region (shown in Fig. 1) [2], [7]–[10]. In contrast, despite having more detailed information in Fundus+Fundus experiments, TokenLearners, which are not guided by explicit anatomical structures, fail to learn features along the vessels (Fig. 6-b). This results in fewer tokens carrying useful features for fovea localization and may limit the effectiveness of the intermediate BTI modules in DSFN, leading to a slight underperformance

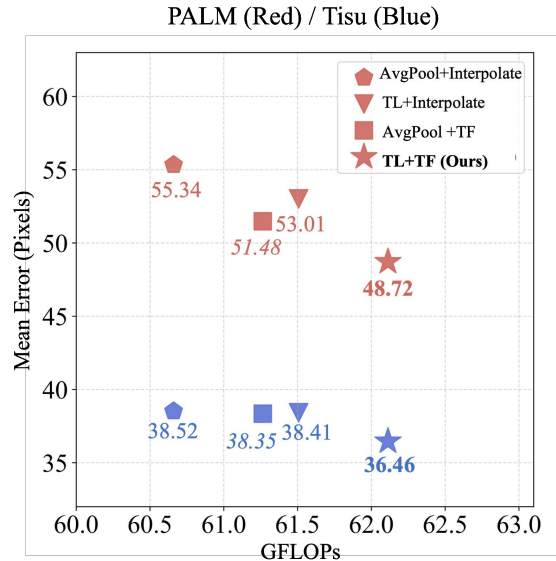


Fig. 7. Visualization of mean errors (Y-axis) of ablation studies for reducing and recovering tokens. X-axis is the computational cost (GFLOPs). The red and blue markers are results on PALM and Tisu datasets, respectively. Numbers below the markers are corresponding mean errors.

TABLE VI
COMPARISON OF PERFORMANCE BETWEEN DIFFERENT CONFIGURATIONS OF LEARNED TOKENS AND LAYER NUMBER OF MHSA BLOCKS UTILIZED IN THE BTI MODULE. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Optimal	Tokens	Layers	Err _d	1/4 R	1/2 R	1R	2R (%)
	8	12	54.40	60	81	93	97
✓	4 × 4	12	52.45	61	78	93	97
	8 × 8	12	48.72	69	85	97	98
	16 × 16	12	52.66	64	84	94	98
	8 × 8	8	49.96	66	82	96	98
	8 × 8	16	55.06	68	85	95	97

on PALM and Tisu datasets (Table V). In the first column of Fig. 6, the predictions of Fundus+Vessel (green crosses) show superior results compared to those of Fundus+Fundus (red crosses). Therefore, the structural information provided by vessel maps as the second input is crucial for achieving accurate fovea localization.

2) *Comparison of Methods for Reducing and Recovering Tokens*: To evaluate the effectiveness of different components within our BTI module for reducing and recovering token numbers, we perform a comprehensive set of ablation experiments on the PALM and Tisu datasets. Instead of employing TokenLearner (TL) and TokenFuser (TF) with adaptively learnable parameters, we alternatively test more straightforward methods used in [29], average pooling (AvgPool) and bilinear interpolation (Interpolate), respectively.

In Fig. 7, we demonstrate a visualization of the mean error against the computational cost (FLOPs) on PALM (red) and Tisu (blue) datasets. Experiments using both TokenLearner and TokenFuser (TL+TF) achieve the best performance on PALM and Tisu with mean errors of 48.72 and 36.46 pixels, respectively. This is in contrast to using more straightforward

methods such as average pooling (AvgPool) and bilinear interpolation (Interpolate), which may lead to a loss of information and reduced performance. In the proposed DSFN (TL+TF), total excess costs of FLOPs for the four BTI modules utilizing TokenLearner and TokenFuser are only **0.85G** and **0.61G**, respectively. This slight increase in computation leverages significant performance benefits, demonstrating the high efficacy of the adaptively learnable parameters of TokenLearner and TokenFuser in our architecture.

3) Comparison of Hyperparameter Settings: The above subsections have evaluated the necessity of vessel distribution inputs and of the BTI module. We also conduct ablation studies to evaluate, in the BTI module, how the performance of feature integration and fusion is affected by the structure of the MHSA blocks, to ensure optimal configurations.

Table VI presents a comparative analysis of the performance on PALM between different configurations of the number of learned tokens and Multi-Head Self-Attention (MHSA) layers utilized in the BTI module. The table evaluates various configurations based on several metrics: mean error, and success rates at different thresholds (1/4 R, 1/2 R, 1R, and 2R). The optimal configuration identified is the use of 8×8 learned tokens with 12 MHSA layers, which achieves the best results across all metrics, including the lowest mean error (48.72) and the highest success rates at all thresholds. Other configurations, such as 4×4 tokens with 12 layers and 8×8 tokens with 8 layers, show lower performance, indicating that both the number of tokens and the depth of MHSA layers significantly influence the performance of feature integration and the accuracy of fovea localization.

VI. CONCLUSIONS

Summary. Accurate detection of the fovea is crucial for diagnosing retinal diseases. While anatomical structures outside the fovea, such as blood vessel distribution, are related to the fovea, few deep learning approaches exploit them to enhance fovea localization performance. In this paper, we propose a novel architecture, DualStreamFoveaNet (DSFN), which effectively fuses features from the retina and corresponding vessel distribution to achieve robust fovea localization. The DSFN comprises a two-stream encoder for multi-cue fusion and a decoder for generating fovea localization maps. Specifically, We introduce the Bilateral Token Incorporation (BTI) with learnable tokens to enhance the efficiency of transformer-based fusion from both fundus and vessel images. Comprehensive experiments demonstrate the advantages of using DSFN, including more accurate localization results through enhanced Multi-Cue Fusion, insensitivity to diseased images, and computational efficiency. Furthermore, our proposed architecture sets a new state-of-the-art on two public datasets (Messidor and PALM) and one large-scale private dataset (Tisu) with metrics ranging from 1/4 R to 2R. Additionally, it outperforms other methods in cross-dataset experiments, showcasing superior generalization capacity and robustness across datasets.

Limitations and Future Works. Although the promising performance of the proposed DSFN shown in our extensive

experiments, the quality of the pseudo-mask of vessel structure still slightly affects the final fovea localization performance, as shown in Table V. Compared to a fuzzy fovea region with an indistinguishable appearance from nearby textures, vessels are very distinct structures with high contrast. In this paper, the significant performance increase of DSFN relies on the vessels' pseudo-masks generated by the TransUNet, which was trained on a small-scale vessel dataset. Recent studies [59]–[61] have specifically focused on vessel structure prediction, potentially providing more accurate pseudo-masks of vessels for our architecture in future research. Moreover, recently emerged Segment Anything Model (SAM)-based segmentation foundation models [62], [63], and their medical fine-tuned versions [64], [65], have shown significant disadvantages in predicting regions with no distinct boundaries and textures. Therefore, our design (*i.e.*, utilizing global vessel structures to explicitly assist fovea localization) remains promising in this area.

Furthermore, the design of DSFN can be utilized as a general architecture for global feature integration and anatomical structure-guided segmentation tasks, as the input of the main and second streams can be any related images. In future work, we may upgrade DSFN for more medical imaging areas. In clinical practice, there are other tasks where even experienced clinicians require anatomical structure-based disease diagnosis. For example, in lung cancer screening, CT scans are commonly used for early detection. With the help of rib structures, doctors can better locate and observe lung lesions. In pelvic MRI, the structure of the pelvis can assist in the detailed evaluation of nearby pelvic organs and aid in diagnosing endometriosis and pelvic tumors.

APPENDIX

A. Notation Summary

All notions and abbreviations of this paper have been summarized in Table VII.

B. Augmentation Settings

For all input images, to reduce overfitting, we adopt a data augmentation strategy utilizing a PyTorch package, albumentations [66]. The details are shown as follows:

RandomResizedCrop (height = 512, width = 512, scale = (0.95, 1.05), ratio = (0.95, 1.05), p = 0.25), HorizontalFlip (p = 0.25), VerticalFlip (p = 0.25), ShiftScaleRotate (shift_limit = 0.0625, scale_limit = 0.05, rotate_limit = 45, p = 0.25), OneOf ([Blur (blur_limit = 5), GaussianBlur (blur_limit = 5), MedianBlur (blur_limit = 5), MotionBlur (blur_limit = 5)], p = 0.25), RandomBrightnessContrast (brightness_limit = 0.1, contrast_limit = 0.1, p = 0.25), OneOf ([CLAHE (clip_limit = 2), IAASharpener(), JpegCompression(), GaussNoise()], p = 0.25), CoarseDropout (p = 0.1).

REFERENCES

- [1] J. Weiter, G. Wing, C. Trempe, and M. Mainster, "Visual acuity related to retinal distance from the fovea in macular disease." *Annals of ophthalmology*, 1984.

TABLE VII
NOTATION SUMMARY

Abbreviation	Explanation
Bi-ViT	Bilateral-ViT: an existing network [15] for fovea localization with bilateral feature fusion design
Bi-ViT/Lit	The light version of Bilateral-ViT
BTI	Bilateral Token Incorporation: The modules designed in this paper for globally bilateral feature fusion with high efficacy
CNN	Convolutional Neural Networks
DSFN	DualStreamFoveaNet: The two-stream architecture proposed in this paper
(G)FLOPs	(Giga) Floating Point Operations
MF	Multi-cue Features
MHSA	Multi-Head Self-Attention
OD	Optic Disc
ROI	Region of Interest
RSU	ReSidual U-blocks: The U-shape-based blocks utilized in the decoder of DSFN to integrate received features
TL and TF	TokenLearner and TokenFuser
ViT	Vision Transformer

Notation	Description
$F_{in}^{main}, F_{in}^{satellite}$	Input tensors from the main and satellite streams, respectively
F_{in}	Concatenated input feature
F_{attn}	Spatial attention map
F'_{in}	Remapped input feature to generate T_L
T_L	Learned tokens
T'_L	Intermediate token representation after MHSA layers
F''_{in}	Remapped input feature to generate F_{out}
F_{out}	Output features generated by T'_L and F''_{in}
$F_{out}^{main}, F_{out}^{satellite}$	Output features equally split by F_{out}
$F_{out}^{main}, F_{out}^{satellite}$	Final output features for the main and satellite streams, respectively

- [2] A. Aquino, "Establishing the macular grading grid by means of fovea centre detection using anatomical-based and visual-based features," *Computers in biology and medicine*, 2014.
- [3] D. Deka, J. P. Medhi, and S. Nirmala, "Detection of macula and fovea for disease analysis in color fundus images," in *International Conference on Recent Trends in Information Systems*, 2015.
- [4] J. P. Medhi and S. Dandapat, "An effective fovea detection and automatic assessment of diabetic maculopathy in color fundus images," *Computers in biology and medicine*, 2016.
- [5] K. S. Deepak and J. Sivaswamy, "Automatic assessment of macular edema from color retinal images," *IEEE Transactions on medical imaging*, vol. 31, no. 3, pp. 766–776, 2011.
- [6] L. Giancardo, F. Meriaudeau, T. P. Karnowski, Y. Li, S. Garg, K. W. Tobin Jr, and E. Chaum, "Exudate-based diabetic macular edema detection in fundus images using publicly available datasets," *Medical image analysis*, vol. 16, no. 1, pp. 216–226, 2012.
- [7] K. M. Asim, A. Basit, and A. Jalil, "Detection and localization of fovea in human retinal fundus images," in *2012 International Conference on Emerging Technologies*, 2012.
- [8] S. Sekhar, W. Al-Nuaimy, and A. K. Nandi, "Automated localisation of optic disk and fovea in retinal fundus images," in *2008 16th European Signal Processing Conference*, 2008.
- [9] H. Li and O. Chutatape, "Automated feature extraction in color retinal images by a model based approach," *IEEE Transactions on biomedical engineering*, 2004.
- [10] H. Narasimha-Iyer, A. Can, B. Roysam, V. Stewart, H. L. Tanenbaum, A. Majerovics, and H. Singh, "Robust detection and classification of longitudinal changes in color retinal fundus images for monitoring diabetic retinopathy," *IEEE transactions on biomedical engineering*, 2006.
- [11] R. GeethaRamani and L. Balasubramanian, "Macula segmentation and fovea localization employing image processing and heuristic based clustering for automated retinal screening," *Computer methods and programs in biomedicine*, 2018.
- [12] X. Guo, H. Wang, X. Lu, X. Hu, S. Che, and Y. Lu, "Robust fovea localization based on symmetry measure," *IEEE Journal of Biomedical and Health Informatics*, vol. 24, no. 8, pp. 2315–2326, 2020.
- [13] S. Pachade, P. Porwal, and M. Kokare, "A novel method to detect fovea from color fundus images," in *Computing, Communication and Signal Processing*. Springer, 2019.
- [14] R. Xie, J. Liu, R. Cao, C. S. Qiu, J. Duan, J. Garibaldi, and G. Qiu, "End-to-end fovea localisation in colour fundus images with a hierarchical deep regression network," *IEEE Transactions on Medical Imaging*, 2020.
- [15] S. Song, K. Dang, Q. Yu, Z. Wang, F. Coenen, J. Su, and X. Ding, "Bilateral-vit for robust fovea localization," in *2022 IEEE 19th International Symposium on Biomedical Imaging*. IEEE, 2022, pp. 1–5.
- [16] M. E. Gegundez-Arias, D. Marin, J. M. Bravo, and A. Suero, "Locating the fovea center position in digital fundus images using thresholding and feature extraction techniques," *Computerized Medical Imaging and Graphics*, 2013.
- [17] A. Giachetti, L. Ballerini, E. Trucco, and P. J. Wilson, "The use of radial symmetry to localize retinal landmarks," *Computerized Medical Imaging and Graphics*, 2013.
- [18] B. Dashtbozorg, J. Zhang, F. Huang, and B. M. ter Haar Romeny, "Automatic optic disc and fovea detection in retinal images using super-elliptical convergence index filters," in *International Conference on Image Analysis and Recognition*, 2016.
- [19] F. Girard, C. Kavalec, S. Grenier, H. B. Tahar, and F. Chieret, "Simultaneous macula detection and optic disc boundary segmentation in retinal fundus images," in *Medical Imaging 2016: Image Processing*, 2016.
- [20] J. M. Molina-Casado, E. J. Carmona, and J. García-Feijóo, "Fast detection of the main anatomical structures in digital retinal images based on intra- and inter-structure relational knowledge," *Computer methods and programs in biomedicine*, 2017.
- [21] L. Zhang, S. Feng, G. Duan, Y. Li, and G. Liu, "Detection of microaneurysms in fundus images based on an attention mechanism," *Genes*, vol. 10, no. 10, p. 817, 2019.
- [22] S. Huang, J. Li, Y. Xiao, N. Shen, and T. Xu, "Rtnet: relation transformer network for diabetic retinopathy multi-lesion segmentation," *IEEE Transactions on Medical Imaging*, vol. 41, no. 6, pp. 1596–1607, 2022.
- [23] H. A. Wibawa, R. Sumiharto, A. Harjoko, and M. B. Sasongko, "The methods of determining temporal direction based on asymmetric information of the optic disc for optimal fovea detection," *Symmetry*, vol. 15, no. 9, p. 1631, 2023.
- [24] S. Sedai, R. Tennakoon, P. Roy, K. Cao, and R. Garnavi, "Multi-stage segmentation of the fovea in retinal fundus images using fully convolutional neural networks," in *2017 IEEE 14th International Symposium on Biomedical Imaging*. IEEE, 2017, pp. 1083–1086.
- [25] B. Al-Bander, W. Al-Nuaimy, B. M. Williams, and Y. Zheng, "Multiscale sequential convolutional neural networks for simultaneous detection of fovea and optic disc," *Biomedical Signal Processing and Control*, 2018.
- [26] M. I. Meyer, A. Galdran, A. M. Mendonça, and A. Campilho, "A pixel-wise distance regression approach for joint retinal optical disc and fovea detection," in *Medical Image Computing and Computer Assisted Intervention–2018: 21st International Conference*. Springer, 2018, pp. 39–47.
- [27] Y. Huang, Z. Zhong, J. Yuan, and X. Tang, "Efficient and robust optic

- disc detection and fovea localization using region proposal network and cascaded network,” *Biomedical Signal Processing and Control*, 2020.
- [28] B. J. Bhatkalkar, S. V. Nayak, S. V. Shenoy, and R. V. Arjunan, “Fundusposnet: A deep learning driven heatmap regression model for the joint localization of optic disc and fovea centers in color fundus images,” *IEEE Access*, vol. 9, pp. 159 071–159 080, 2021.
- [29] A. Prakash, K. Chitta, and A. Geiger, “Multi-modal fusion transformer for end-to-end autonomous driving,” in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 7077–7087.
- [30] Y. Guo, D. Zhou, P. Li, C. Li, and J. Cao, “Context-aware poly (a) signal prediction model via deep spatial-temporal neural networks,” *IEEE Transactions on Neural Networks and Learning Systems*, 2022.
- [31] S. Yang, D. Zhou, J. Cao, and Y. Guo, “Lightnet: An integrated learning method for low-light image enhancement,” *IEEE Transactions on Computational Imaging*, vol. 9, pp. 29–42, 2023.
- [32] X. Wang, Z. Guan, W. Qian, J. Cao, C. Wang, and R. Ma, “Stfuse: Infrared and visible image fusion via semisupervised transfer learning,” *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
- [33] Y. Guo, D. Zhou, X. Ruan, and J. Cao, “Variational gated autoencoder-based feature extraction model for inferring disease-mirna associations based on multiview features,” *Neural Networks*, vol. 165, pp. 491–505, 2023.
- [34] W. Li, Y. Guo, B. Wang, and B. Yang, “Learning spatiotemporal embedding with gated convolutional recurrent networks for translation initiation site prediction,” *Pattern Recognition*, vol. 136, p. 109234, 2023.
- [35] M. Ryoo, A. Piergiovanni, A. Arnab, M. Dehghani, and A. Angelova, “Tokenlearner: Adaptive space-time tokenization for videos,” *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 786–12 797, 2021.
- [36] Y. Fu, G. Zhang, J. Li, D. Pan, Y. Wang, and D. Zhang, “Fovea localization by blood vessel vector in abnormal fundus images,” *Pattern Recognition*, vol. 129, p. 108711, 2022.
- [37] O. Ronneberger, P. Fischer, and T. Brox, “U-net: Convolutional networks for biomedical image segmentation,” in *Medical image computing and computer-assisted intervention–2015: 18th international conference*. Springer, 2015, pp. 234–241.
- [38] K. Suzuki, “Overview of deep learning in medical imaging,” *Radiological physics and technology*, vol. 10, no. 3, pp. 257–273, 2017.
- [39] Y. Zhang, J. Shi, Y. Peng, Z. Zhao, Q. Zheng, Z. Wang, K. Liu, S. Jiao, K. Qiu, Z. Zhou *et al.*, “Artificial intelligence-enabled screening for diabetic retinopathy: a real-world, multicenter and prospective study,” *BMJ Open Diabetes Research and Care*, vol. 8, no. 1, p. e001596, 2020.
- [40] G. Murtaza, L. Shuib, A. W. Abdul Wahab, G. Mujtaba, H. F. Nweke, M. A. Al-garadi, F. Zulfikar, G. Raza, and N. A. Azmi, “Deep learning-based breast cancer classification through medical imaging modalities: state of the art and research challenges,” *Artificial Intelligence Review*, vol. 53, no. 3, pp. 1655–1720, 2020.
- [41] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, “U2-net: Going deeper with nested u-structure for salient object detection,” *Pattern Recognition*, vol. 106, 2020.
- [42] Q. Yu, K. Dang, N. Tajbakhsh, D. Terzopoulos, and X. Ding, “A location-sensitive local prototype network for few-shot medical image segmentation,” in *2021 IEEE 18th international symposium on biomedical imaging*. IEEE, 2021, pp. 262–266.
- [43] R. Yang and Y. Yu, “Artificial convolutional neural network in object detection and semantic segmentation for medical imaging analysis,” *Frontiers in oncology*, vol. 11, p. 638182, 2021.
- [44] J. Cheng, L. Dong, and M. Lapata, “Long short-term memory-networks for machine reading,” in *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Nov 2016, pp. 551–561.
- [45] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena, “Self-attention generative adversarial networks,” in *International conference on machine learning*. PMLR, 2019, pp. 7354–7363.
- [46] J. H. Tan, U. R. Acharya, S. V. Bhandary, K. C. Chua, and S. Sivaprasad, “Segmentation of optic disc, fovea and retinal vasculature using a single convolutional neural network,” *Journal of Computational Science*, vol. 20, pp. 70–79, 2017.
- [47] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, “Attention is all you need,” *Advances in neural information processing systems*, vol. 30, 2017.
- [48] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, “An image is worth 16x16 words: Trans-
formers for image recognition at scale,” in *9th International Conference on Learning Representations*, 2021.
- [49] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, “Transunet: Transformers make strong encoders for medical image segmentation,” *CoRR*, vol. abs/2102.04306, 2021.
- [50] E. Decencière, X. Zhang, G. Cazuguel, B. Lay, B. Cochener, C. Trone, P. Gain, R. Ordonez, P. Massin, A. Erginay *et al.*, “Feedback on a publicly distributed image database: the messidor database,” *Image Analysis & Stereology*, 2014.
- [51] H. Fu, F. Li, J. I. Orlando, H. Bogunović, X. Sun, J. Liao, Y. Xu, S. Zhang, and X. Zhang, “Palm: Pathologic myopia challenge,” 2019. [Online]. Available: <https://dx.doi.org/10.21227/55pk-8z03>
- [52] J. Staal, M. D. Abramoff, M. Niemeijer, M. A. Viergever, and B. Van Ginneken, “Ridge-based vessel segmentation in color images of the retina,” *IEEE transactions on medical imaging*, 2004.
- [53] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, “Pyramid vision transformer: A versatile backbone for dense prediction without convolutions,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 568–578.
- [54] J. Park, S. Woo, J.-Y. Lee, and I.-S. Kweon, “Bam: Bottleneck attention module,” in *British Machine Vision Conference*, 2018, pp. 147–161.
- [55] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, “Cbam: Convolutional block attention module,” in *Proceedings of the European conference on computer vision*, 2018, pp. 3–19.
- [56] D. P. Kingma and J. Ba, “Adam: A method for stochastic optimization,” in *International Conference on Learning Representations (Poster)*, 2015.
- [57] J. Sigut, O. Nuñez, F. Fumero, S. Alayon, and T. Diaz-Aleman, “Fovea localization in retinal images using spatial color histograms,” *Multimedia Tools and Applications*, vol. 83, no. 6, pp. 17 753–17 771, 2024.
- [58] I. Sobh, L. Amin, S. Abdelkarim, K. Elmawdy, M. Saeed, O. Abdelawab, M. Gamal, and A. El Sallab, “End-to-end multi-modal sensors fusion system for urban automated driving,” in *Advances in Neural Information Processing Systems Workshops*, 2018.
- [59] M. Jiang, Y. Zhu, and X. Zhang, “Covi-net: A hybrid convolutional and vision transformer neural network for retinal vessel segmentation,” *Computers in Biology and Medicine*, vol. 170, p. 108047, 2024.
- [60] W. Ding, Y. Sun, J. Huang, H. Ju, C. Zhang, G. Yang, and C.-T. Lin, “Rcar-unet: Retinal vessel segmentation network algorithm via novel rough attention mechanism,” *Information Sciences*, vol. 657, p. 120007, 2024.
- [61] Y. Tan, K.-F. Yang, S.-X. Zhao, J. Wang, L. Liu, and Y.-J. Li, “Deep matched filtering for retinal vessel segmentation,” *Knowledge-Based Systems*, vol. 283, p. 111185, 2024.
- [62] A. Kirillov, E. Mintun, N. Ravi, H. Mao, C. Rolland, L. Gustafson, T. Xiao, S. Whitehead, A. C. Berg, W.-Y. Lo *et al.*, “Segment anything,” in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 4015–4026.
- [63] L. Ke, M. Ye, M. Danelljan, Y.-W. Tai, C.-K. Tang, F. Yu *et al.*, “Segment anything in high quality,” *Advances in Neural Information Processing Systems*, vol. 36, 2024.
- [64] J. Cheng, J. Ye, Z. Deng, J. Chen, T. Li, H. Wang, Y. Su, Z. Huang, J. Chen, L. Jiang *et al.*, “Sam-med2d,” *arXiv preprint arXiv:2308.16184*, 2023.
- [65] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, “Segment anything in medical images,” *Nature Communications*, vol. 15, no. 1, p. 654, 2024.
- [66] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, “Albumentations: Fast and flexible image augmentations,” *Information*, vol. 11, no. 2, 2020.