RVD: A Handheld Device-Based Fundus Video Dataset for Retinal Vessel Segmentation

MD Wahiduzzaman Khan[1*] **Hongwei Sheng**[1 2*] **Hu Zhang**[2*] **Heming Du**[3]
Sen Wang[2] **Minas Theodore Coroneo**[4] **Farshid Hajati**[5] **Sahar Shariflou**[1]
Michael Kalloniatis[6] **Jack Phu**[4] **Ashish Agar**[4] **Zi Huang**[2]
Mojtaba Golzan[1†] **Xin Yu**[2†]
[1]University of Technology Sydney [2]University of Queensland [3]Australian National University
[4]University of New South Wales [5]Victoria University [6]University of Houston-Downtown
mdwahiduzzaman.khan@student.uts.edu.au
[*] Equal Contribution [†] Corresponding Author

Abstract

Retinal vessel segmentation is generally grounded in image-based datasets collected with bench-top devices. The static images naturally lose the dynamic characteristics of retina fluctuation, resulting in diminished dataset richness, and the usage of bench-top devices further restricts dataset scalability due to its limited accessibility. Considering these limitations, we introduce the first video-based retinal dataset by employing handheld devices for data acquisition. The dataset comprises 635 smartphone-based fundus videos collected from four different clinics, involving 415 patients from 50 to 75 years old. It delivers comprehensive and precise annotations of retinal structures in both spatial and temporal dimensions, aiming to advance the landscape of vasculature segmentation. Specifically, the dataset provides three levels of spatial annotations: binary vessel masks for overall retinal structure delineation, general vein-artery masks for distinguishing the vein and artery, and fine-grained vein-artery masks for further characterizing the granularities of each artery and vein. In addition, the dataset offers temporal annotations that capture the vessel pulsation characteristics, assisting in detecting ocular diseases that require fine-grained recognition of hemodynamic fluctuation. In application, our dataset exhibits a significant domain shift with respect to data captured by bench-top devices, thus posing great challenges to existing methods. Thanks to rich annotations and data scales, our dataset potentially paves the path for more advanced retinal analysis and accurate disease diagnosis. In the experiments, we provide evaluation metrics and benchmark results on our dataset, reflecting both the potential and challenges it offers for vessel segmentation tasks. We hope this challenging dataset would significantly contribute to the development of eye disease diagnosis and early prevention. The dataset is available at RVD.

1 Introduction

The observation of the retinal vasculature patterns serves as a reliable approach to tracking the morphological changes of eyes over time. These morphological changes have been found to be closely associated with a spectrum of ocular diseases, *e.g.*, diabetic retinopathy, age-related macular degeneration, and glaucoma [83, 14]. Retinal vessel segmentation aims to provide pixel-level extraction of the visible vasculature from a fundus image [56]. It is the initial yet fundamental step in objectively assessing vasculature in fundus images and quantitatively interpreting associated morphometrics. Thus, this task plays a pivotal role in understanding and diagnosing ocular diseases.

37th Conference on Neural Information Processing Systems (NeurIPS 2023) Track on Datasets and Benchmarks.

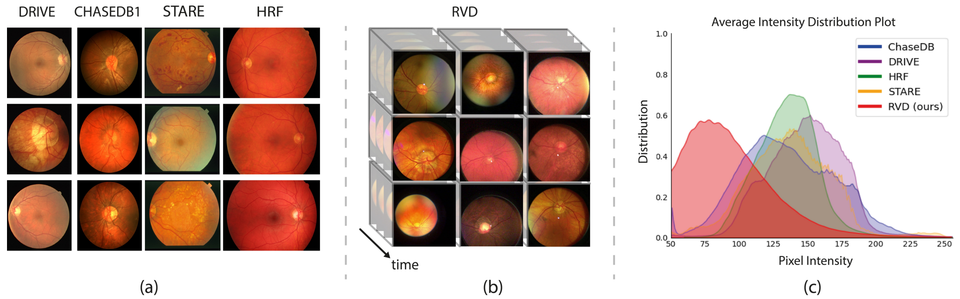


Figure 1: **(a)** Samples from existing image based retinal vessel datasets: DRIVE [72], STARE [69], HRF [38], and CHASE_DB1 [25]. **(b)** Video samples from our retinal vessel dataset. Different from existing image-based datasets, our dataset captures continuous changes in retinal vessels and facilitates the analysis of vessel dynamics in the retina. **(c)** The intensity distributions of our dataset and existing ones. The differences imply the domain gaps between our dataset and existing ones.

Existing methods for retinal vessel segmentation are designed on image-based datasets [38, 72, 69, 25], as shown in Fig. 1 (a). Although these datasets have contributed valuable vessel annotations for studying retinal segmentation, the static nature of images inherently limits their ability to portray dynamic retinal characteristics, *e.g.*, vessel pulsations. These dynamic phenomena play a vital role in facilitating comprehensive and in-depth understanding of retinal functionality and vasculature structure. Moreover, image-based datasets are captured by expensive bench-top ophthalmic equipment, which is operated by professionally trained clinicians [31, 34]. Such requirements potentially limit the scale of the datasets and data diversity, thereby adversely affecting the generalization ability of the models trained on these datasets.

In recent years, advances in imaging technology have enabled the usage of smartphone-based devices for retinal observation [79, 32]. They offer better flexibility and portability, allowing for scalable data collection. In this paper, we introduce the first video-based retinal vessel dataset (RVD), a collection of 635 smartphone-based videos with detailed vessel annotation. These videos are recorded from four clinics, including patients from 50 to 75 years old. Some examples of our dataset are shown in Fig. 1 (b). The sequential frames capture the continuous changes in retinal vessels and thus significantly facilitate the analysis of subtle fluctuations in the retinal structure. Therefore, the use of portable devices for data acquisition and the provision of the video modality remarkably overcome the limitations of existing datasets.

The annotations provided in our dataset span two dimensions: spatial and temporal. In the spatial dimension, we offer three distinct levels of annotations: binary vessel masks, general vein-artery masks, and fine-grained vein-artery masks. Each kind of annotation is tailored to specific clinical purposes. Specifically, for the binary vessel masks, we identify the sharpest and most representative frame from the video clip and generate binary masks representing the skeletal structure of the vessels. This mask primarily targets the holistic vessel structure but neglects the difference between arteries and veins. For general vein-artery masks, we differentiate veins and arteries based on their respective vessel calibres and generate separate masks for them respectively. Lastly, in contrast to the general differentiation between arteries and veins, the fine-grained vein-artery masks further divide each retinal artery and vein into sections based on a set of pre-defined vessel widths. We thus generate eight different vein-artery masks for each sample and these masks precisely reflect the granularities of retinal vessels. These sophisticated masks are highly demanded when detecting ocular diseases [3, 9].

In the temporal dimension, we enrich our dataset with annotations of the complex dynamics of retinal vasculature. For each video, we focus on the optic disk regions where the retinal vessel fluctuation normally occurs. We then select and annotate frames with the maximal and minimal pulse widths as well as label the existence of spontaneous retinal venous pulsations (SVP). The existence and extent of vessel changes signify vascular pulsations and cranial pressure-related alterations. Clinically, the signals of pulsation facilitate the detection of abnormalities in retinal vessels, while precise identification of pressure-related alterations aids in detecting temporally-dependent ocular diseases. Our integration of temporal annotations thus increases its potential for ocular disease diagnosis.

The distinction between smartphone-based and benchtop devices and data modality differences result in domain gaps, as illustrated in Fig. 1. Furthermore, since our data are collected by handheld devices in clinics, our dataset also involves more realistic factors, *e.g.*, the operations of the clinicians, surrounding illumination conditions, and eye movements of patients during video capture. Consequently, our dataset presents more challenges for existing vessel segmentation methods. More importantly, the large number of training samples and detailed annotations in our dataset will likely pave the way toward more advanced yet portable retinal analysis and more accurate disease diagnosis. In the experiments, we delve into an in-depth analysis of our dataset and provide benchmark results of different tasks on our newly curated dataset.

The main contributions of our paper are summarized as follows:

- **Dataset construction**: We construct a new video-based retinal vessel dataset (RVD) with rich spatial and temporal annotations for vessel segmentation tasks. To the best of our knowledge, RVD is the first mobile-device based dataset for retinal vessel segmentation.
- **Three-level spatial annotations**: Our dataset introduces three levels of annotations in spatial, comprising binary vessel masks, general vein-artery masks, and fine-grained vein-artery masks. The hierarchical and diverse descriptions of spatial annotations enable us to better analyze the vessel structure.
- **Temporal annotations**: Our dataset also provides temporal annotations of spontaneous retinal venous pulsations (SVP) to reveal the dynamic changes in retinal vessels. This enables the assessment of pulsatile variations in retinal vessels.
- **Benchmarking**: We investigate the gap between our dataset and previous retinal datasets by assessing the performance of several state-of-the-art methods. The experimental results will shed some light on mobile-device based retinal vessel segmentation.

2 Related Work

Existing Retinal Datasets: In the realm of retinal vessel segmentation, various retinal vessel datasets have been proposed. Existing datasets can be roughly categorized into two streams: binary vessel based ones and artery-vein based ones. Among the datasets with binary vessel masks, DRIVE [72], STARE [69], HRF [38], and CHASE_DB1 [25] have emerged as the most frequently used datasets. In fact, each of these datasets only comprises dozens of images. For example, DRIVE consists of 40 images captured in the 45-degree field of view, with an image size of 584×565 pixels. Besides, DRiDB [59], ARIA [23], IOSTAR [84], and RC-SLO [1] are another publicly available datasets for retinal vessel segmentation. However, they are less used in recent years considering their data quality and maintenance. Recently, the FIVES dataset has been introduced [34], with data distributed across four categories: normal retinas, retinas affected by Diabetic Retinopathy, Glaucoma, and Age-related Macular Degeneration. It comprises 800 retinal images.

Regarding the datasets with artery-vein masks, RITE [31], AV-DRIVE [60], INSPIRE-AVR [57], and WIDE [21] are the available ones. The AV-DRIVE dataset, derived from DRIVE, consists of 40 images and offers separate ground truth masks for arteries and veins. The INSPIRE-AVR is an independently constructed dataset with artery-vein ground truth masks. It consists of 40 color images in total. The WIDE dataset provides 30 scanning laser ophthalmoscope (SLO) images.

In contrast to existing datasets which are collected with cumbersome bench-top devices and are composed of static images, our dataset is constructed with portable handheld devices and is video-based. Our dataset preserves the dynamic characteristics of vessels. Besides, existing datasets typically provide only one type of annotation for a specific research purpose, whereas our dataset offers annotations in both spatial and temporal dimensions. The spatial annotations include binary vessel masks, general vein-artery masks, and fine-grained vein-artery masks, respectively. The temporal annotations reveal the state of SVP, an important signal for diagnosing various diseases.

Methods for Retinal Vessel Segmentation: In the past years, a variety of methods have been developed for retinal vessel segmentation. Traditional methods mainly depend on handcrafted features [6, 16, 88, 76, 61, 28, 17, 36], which are less discriminative and effective [51]. With the unprecedented breakthroughs of deep neural networks (DNNs) in the image classification, detection, and segmentation tasks, researchers have explored the potential of DNNs in retinal vessel segmentation [22, 43, 35, 81, 89]. Many works [5, 42, 71, 58, 33, 15] adopt fully convolutional networks [46] to produce more accurate segmentation of retinal images by combining semantic information from deep layers with appearance information from shallow layers. Several works

Table 1: Comparisons of different retinal vessel segmentation datasets. “Num” denotes the number of annotated image frames.

Dataset	Resolution	Modality	Device	Num	Dimension	Annotation type
STARE [69]	605×700	Image	Benchtop	20	Spatial	Binary
DRIVE [72]	768×584	Image	Benchtop	40	Spatial	Binary
ARIA [23]	576×768	Image	Benchtop	161	Spatial	Binary
CHASEDB1 [25]	990×960	Image	Benchtop	28	Spatial	Binary
INSPIRE-AVR [57]	2392×2048	Image	Benchtop	40	Spatial	Multi-class
HRF [38]	3304×2336	Image	Benchtop	45	Spatial	Binary
RITE [31]	768×584	Image	Benchtop	40	Spatial	Multi-class
FIVES [34]	2048×2048	Image	Benchtop	800	Spatial	Binary
RAVIR [29]	768×768	Image	IR Laser	42	Spatial	Multi-class
RVD (ours)	**1800×1800**	**Video**	**Hand-held**	**1,270**	**Spatial + Temporal**	**Multi-class**

have focused on modifying the U-Net structure [74, 64, 85] for vessel segmentation. [86] first introduces the residual connection into U-net to detect vessels. This idea has been adopted in later studies [4, 41, 26, 78]. [82] introduces the local-region and cross-dataset contrastive learning losses in training to explore a more powerful feature embedding space. Besides, several other methods employ various networks and strategies for retinal vessel segmentation, such as generative and adversarial networks [40, 70], ensemble learning [49, 75, 44], and graph convolutional network [68].

The aforementioned methods are mainly conducted on datasets DRIVE, STARE, CHASE_DB1, and HRF, with binary vessel masks as supervision. Thanks to INSPIRE-AVR, AV-DRIVE, and WIDE, many works have been proposed to distinguish artery and vein [18, 21, 87, 50]. In [47], a multi-task deep neural network with spatial activation is proposed. The constructed network is able to segment full retinal vessels, arteries, and veins simultaneously. More recently, transformer based models, *e.g.*, ViT [20], Swin transformer [45], and Mask2Former [13], have been proposed. These models have demonstrated their superior performance in capturing visual concepts and become popular backbones in visual understanding tasks [54]. We thus choose these models in the experiments to study the characteristics of our proposed dataset.

3 Our Proposed RVD

In this section, we first describe our data collection process and data sources[1] Concerning privacy and ethic, we perform this study in accordance with the guidelines of the Tenets of Helsinki. Written consent was obtained from all participants prior to any data collection, and all examination protocols adhered to the tenets of the Declaration of Helsinki. Once clinical data have been collected, we need to clean and pre-process the data in order to facilitate clinicians’ annotations and neural network training. In our work, the annotations are provided by professionally well-trained clinicians, and they have been asked to not only annotate conventional spatial segmentation masks but also temporal segmentation masks for dynamic biomarkers, such as Spontaneous retinal Venous Pulsations (SVPs). Last, we will introduce the data split and relevant tasks that are supported by our dataset.

3.1 Data Collection

For data collection, the employed hand-held fundus imaging devices are constructed by connecting a smartphone to the fundus camera lens. Then, clinicians are trained to operate the hand-held devices to examine patients’ retinas while collecting fundus videos. These participants are fully aware of data collection when they undergo their annual medical examinations. With the help of clinicians, a total of 415 patients from four different clinics participate in the data collection process. As data are collected in different clinics over the past five years, the employed smartphones are different, thus increasing the diversity of data sources. More specifically, 264 males and 151 females are included here. Their ages range from 50 to 75. People of these ages are commonly considered to be at high risk for eye-related diseases, such as glaucoma and hypertension [37]. Our dataset involves both videos recorded from healthy eyes and videos from eyes with ocular diseases.

During the collection, one eye of each patient is recorded at a time. In this manner, at least one fundus video of each patient has been recorded and some participate multiple times in video recording. As a

[1] Our dataset follows the copyright **Creative Commons BY-NC-ND 4.0** license ©.

result, a total of 635 RGB videos have been captured. All captured videos have a frame rate of 25 frames per second, with the duration varying between 2 to 30 seconds. The total number of frames in our dataset is over 130,000. This collection process ensures the generality and diversity of our dataset for retinal vessel analysis. The detailed information of our dataset is shown in Table 1 and some examples could be found in Fig. 1 (b).

3.2 Data Cleaning and Preprocessing

Although we have tried our best to minimize environmental interference during collection, the original videos still exhibit various noise, such as video jittering and motion blur. Such noise will severely degrade the quality of collected videos and impose more difficulties in annotations. Hence, we eliminate the noise in the footage to improve the quality of our dataset and facilitate annotations.

Data cleaning: Considering that blood vessel dynamics mostly appear in ODR, we remove the video segments without ODR and ensure the existence of ODR in all videos. To this end, we employ an ODR detection method to localize ODR. Specifically, we label ODR regions by bounding-boxes, and for each video, we only annotate one frame per 25 frames (*i.e.*, 1 second), similar to [67]. Then, we leverage the labeled ODR as supervision to train the Faster-RCNN detection network [62]. After that, the remaining frames are labeled by the trained Faster-RCNN. Manual check is also conducted to modify erroneous detection results by annotators. We only select video segments in the dataset if their ODRs are detectable in a minimum of 30 continuous frames. Such operations help maintain the overall quality of our video data.

To further improve the data quality, we leverage the optical flow to pinpoint frames with a high level of blur. Optical flow captures the spatial alterations between distinct frames, and thus it could serve as an indicator of spatial sharpness. Frames with large optical flow are subsequently discarded, as they likely correspond to instances of blurring. Similar to ODR detection, annotators (non-experts) also manually remove frames that undergo severe blur but have not been spotted by optical flow.

Data Preprocessing: In retinal vessel segmentation, ODR and its surrounding area are the most representative regions of the eye and provide extensive details about retinal vessels. However, the inherent ocular movement results in varying ODR positions across different frames. Such variations can impede the precise observation and annotation of SVP by clinicians. To tackle this issue, we employ the template matching algorithm [7] to stabilize the ODRs across video segments, ensuring a consistent ODR placement and a fixed field of view across frames. This facilitates human observation and machine perception of dynamic changes surrounding the ODR, thus greatly enhancing annotations and clinical diagnosis.

3.3 Data Annotation

To ensure the annotation quality, six clinicians are involved in this process. In the process of annotation, we also incorporate inter-rater reliability using the IoU metric to ensure the consistency of annotations across different annotators. Specifically, after every 100 annotations, we will randomly select 5 frames from each clinician and then reassign the frames to another clinician for annotations. If the IoU is lower than 0.95, the annotations will be manually reviewed by a third clinician. For the temporal SVP annotations, each video has been annotated by three clinicians to reduce annotation bias.

3.3.1 Spatial segmentation

To mitigate the redundancy of annotating similar video frames, we currently select and annotate two frames for each video. Here, we adopt a three-fold strategy to identify the most representative frame from a video. **(1)** A frame is in high image quality (high sharpness/contrast) and contains the most vessels in a video; **(2)** A frame can cover ODR, fovea, and macula regions without obstructions. These regions pathologically have a significant number of capillaries; **(3)** Based on the first frame, we choose another frame that not only has visible ODR with high-density vessels but also has a maximum spatial distance between its ODR and the previously selected ones. In this way, the annotated two frames will cover most of the retinal vasculature and any pathological regions of the retina. We then elaborate on creating three types of spatial annotations: binary vessel masks, general vein-artery masks, and fine-grained vein-artery masks, as follows:

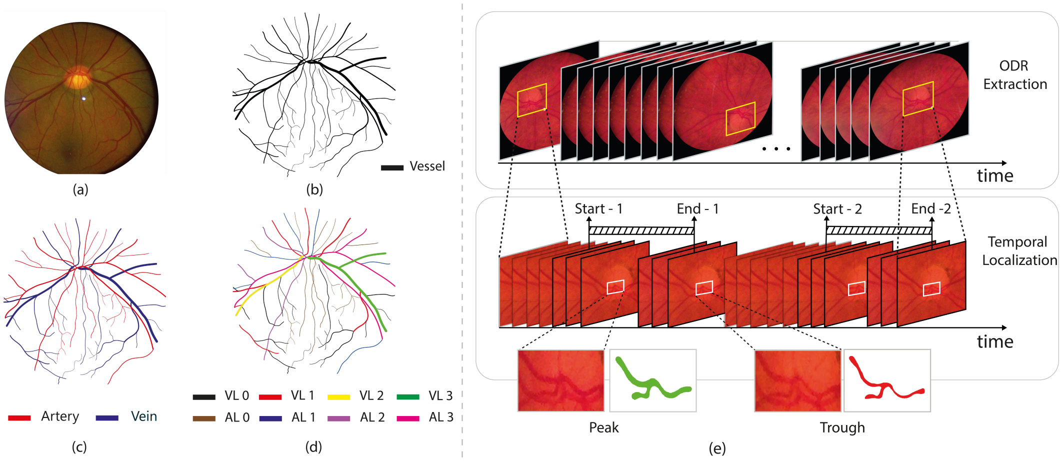


Figure 2: **Left:** Illustration of our multi-grained segmentation annotations. For each given fundus image **(a)**, we provide three different kinds of segmentation masks including a conventional binary mask **(b)**, a general artery-vein mask **(c)** and a fine-grained artery-vein mask **(d)** (VL: vein width level, AL: artery width level, the numbers (0 to 3) indicate four increasing width levels). **Right**: Overview of the temporal annotations **(e)**, including ODR locations, presence and absence of SVP, temporal localization of SVP, and "peak" and "trough" of SVP.

Binary vessel masks: To generate the binary vessel masks, we adopt a similar method proposed in [48]. For each frame, we first draft a centerline-level annotation using the ImageJ software [65] and generate the delineation of vessel boundaries to obtain the main structure of vessels. Then we employ our experts to manually refine the structure by correcting the boundaries and improving the details of small capillaries. We can obtain the binary vessel masks by assigning the label to the refined structure (see Fig. 2 (b)).

General vein-artery masks: Many intracranial vascular diseases are found to be related to retinal vessels and affect the arteries and veins differently [2]. Thus, distinguishing between the retinal artery and vein plays a critical role in the clinical biomarker study of how various systemic and cardiovascular diseases affect the retinal vessels. In practice, the arteries and veins can be distinguished based on their difference in three aspects: color, light reflection, and calibres. The veins generally have a darker color than arteries and show a smaller central light reflex. Meanwhile, the veins are also wider than adjacent arteries. Then, clinicians only need to assign labels (*i.e.*, vein and artery) to the vessels and obtain the vein-artery masks, as shown in Fig. 2 (c).

Fine-grained vein-artery masks: Vascular morphology holds substantial clinical significance, as alterations in vessel diameters frequently signify the presence of various diseases. For example, damage of the small retinal vessels could result in diabetic retinopathy [39]. Similarly, glaucoma pathogenesis is postulated to be linked to alterations in the retinal vasculature, such as retinal arteriolar narrowing and decreased fractal dimension [10]. Despite the clinical importance of such information, existing datasets scarcely provide this type of labels. Therefore, we consider the morphological characteristics of each artery and vein in our dataset and thus provide fine-grained vein-artery masks based on the vessel widths.

Specifically, we first measure the vessel diameters automatically via "Vessel Diameters" plugin in ImageJ.[2] Then, we divide the arteries into multiple small vessel segments based on the diameters of the vessels. Based on the largest diameter among these artery segments, we define four levels of widths according to specific ratios. More specifically, a vessel segment within the range of 0-25% of the largest diameter is categorized as level 0. Similarly, levels 1, 2, and 3 correspond to vessel widths in the ranges of 25%-50%, 50%-75%, and 75%-100% of the largest diameter, respectively. Afterward, we obtain four-class masks for arteries based on vessel widths. The same operation is also applied to veins. After the automatic processing, clinicians will validate the quality of the fine-grained

[2] https://imagej.net/software/imagej/

segmentation masks. This process ultimately yields eight-class masks for both arteries and veins (see Fig. 2 (d)). Those masks significantly enrich the granularity of our dataset.

3.3.2 Temporal localization

Existence of SVP: Based on the results of data cleaning and preprocessing, we utilize the stabilized videos and further annotate the dynamic state of vessel pulsations. Spontaneous retinal Venous Pulsation (SVP) plays a crucial role as a biomarker in retina assessments. Specifically, SVP is characterized by rhythmic pulsations evident in the central retinal vein and its branches, typically observable within the optic disc region (ODR) of the retinas. The absence of SVP holds substantial clinical significance, as it is correlated with certain pathologies. For example, the absence of SVP is associated with progressive glaucoma [53], and it is indicative of increased intracranial pressure [52]. Considering the requirement of specialized knowledge, we have invited multiple clinicians and finished the annotation of SVP presence or absence for each video in our dataset. Once the annotation process completes, we obtain 335 "SVP-present" videos and 300 "SVP-absent" videos respectively. This annotation establishes a fundamental task of SVP detection, facilitating further analysis and investigation on the relationship between SVP and eye diseases.

Temporal duration of SVP: After annotating the existence of SVP in the stabilized videos, some "SVP-present" videos may not contain SVP throughout the whole video. This means in some frames SVP is not visible. Using these videos to train an SVP classification model would suffer ambiguity especially when an entire video cannot be fed into a neural network. Therefore, we further provide temporal emergence annotations of SVP by indicating the starting and ending frames of retinal vessel fluctuation (see Fig. 2 (e)). The detailed duration of SVP serves two purposes: it acts as a valuable signal to improve the performance of SVP detection tasks and concurrently sets a new task for SVP temporal localization. We obtain videos in three distinct groups: 156 videos containing intermittent SVPs, 179 videos demonstrating persistent SVPs, and the remaining 300 videos without SVPs. These temporal annotations allow us to better understand retinal vessel dynamics.

"Peak" and "Trough" annotations of SVP: As discussed above, SVP reflects the temporal dilation and contraction in retinal vessels. The state with maximal dilation is characterized as "peak", whereas those with maximal contraction are termed "trough". Here, we select frames corresponding to the "peak" and "trough" states from each "SVP-present" video. Subsequently, we generate corresponding masks for these selected frames, yielding a total of 670 annotated masks. This annotation allows us to quantitatively measure the extent of pulsations and occurring positions of vessel pulsations.

3.4 Data Protocols

Data split: When partitioning our dataset for training and evaluation, we also take into account the similarity of the recorded videos of the same person. Specifically, we ensure that videos captured from the same patient are allocated to the same subset during the partitioning process. This strategy aims to decrease the similarity between training data and testing data and thus minimizes performance bias. In practice, we divide the data based on patient IDs. In this manner, the same patient's videos will not appear simultaneously in both the training and testing sets. Then, we select 517 videos from the 635 videos for training and validation, and the rest 118 videos are used for testing. We also cross-validate a method with three different data splits and will release the dataset and data splits.

Metrics: Based on our annotations, we can conduct tasks in two major categories. (1) Retinal vessel segmentation metrics: Since the binary, general artery-vein and fine-grained vessel segmentation tasks are essentially semantic segmentation, we adopt the mean Intersection over Union (mIoU), mean Accuracy (mAcc), and macro-averaged F1-score (mFscore) to evaluate the performance of models on our dataset. Note that, mFscore is the average of F1-score by treating all classes equally and the F1 score is also known as the DICE score in binary classification or segmentation tasks[3], the mFscore thus essentially equals the mean of DICE scores for segmentation tasks. (2) SVP recognition and temporal localization metrics: SVP recognition is to classify whether SVP exists in a video. SVP localization task is to identify the time period where SVP appears in a video. We adopt the Accuracy (Acc), Area Under the Receiver Operating Characteristic Curve (AUROC), and Recall for SVP recognition. The frame-mAP (F-mAP), video-mAP (V-mAP) [27] under IoU threshold 0.5 and mean Intersection over Union (mIOU) are adopted for the task of SVP localization.

[3] https://en.wikipedia.org/wiki/F-score

Table 2: Segmentation results of different methods on our RVD dataset. "DLV3" denotes "DeepLabV3" and "M2F" denotes "Mask2Former".

Method	Backbone	Binary			General Artery-Vein			Fine-grained Artery-Vein		
		mIoU	mAcc	mFscore	mIoU	mAcc	mFscore	mIoU	mAcc	mFscore
DLV3 [11]	UNet [63]	$66.59_{\pm 0.6}$	$72.92_{\pm 1.0}$	$76.67_{\pm 0.9}$	$36.54_{\pm 1.5}$	$37.80_{\pm 1.9}$	$55.85_{\pm 1.2}$	$12.81_{\pm 0.4}$	$14.03_{\pm 0.7}$	$19.16_{\pm 0.5}$
	ResNet50 [30]	$62.15_{\pm 0.6}$	$65.84_{\pm 1.1}$	$70.72_{\pm 0.7}$	$47.92_{\pm 0.1}$	$51.85_{\pm 0.2}$	$57.21_{\pm 0.1}$	$17.30_{\pm 0.3}$	$19.59_{\pm 0.7}$	$24.67_{\pm 0.5}$
	ResNet101	$62.89_{\pm 0.7}$	$71.45_{\pm 1.0}$	$78.45_{\pm 1.3}$	$56.60_{\pm 0.7}$	$51.99_{\pm 0.9}$	$57.10_{\pm 1.0}$	$18.13_{\pm 0.5}$	$21.05_{\pm 0.6}$	$24.72_{\pm 0.6}$
M2F [12]	ResNet50	$70.27_{\pm 0.3}$	$77.65_{\pm 0.2}$	$79.51_{\pm 0.3}$	$57.60_{\pm 0.1}$	$66.80_{\pm 0.2}$	$69.06_{\pm 0.1}$	$24.88_{\pm 0.8}$	$32.58_{\pm 1.5}$	$34.11_{\pm 1.1}$
	ResNet101	$70.74_{\pm 0.3}$	$78.78_{\pm 0.3}$	$79.99_{\pm 0.3}$	$59.43_{\pm 1.7}$	$68.58_{\pm 1.5}$	$70.73_{\pm 1.7}$	$31.62_{\pm 4.0}$	$41.96_{\pm 5.1}$	$42.89_{\pm 6.4}$
	Swin-T [45]	$70.94_{\pm 0.7}$	$78.87_{\pm 0.7}$	$80.14_{\pm 0.7}$	$58.58_{\pm 1.2}$	$69.10_{\pm 2.6}$	$71.39_{\pm 0.1}$	$28.14_{\pm 3.2}$	$36.93_{\pm 4.3}$	$38.36_{\pm 4.2}$
	Swin-S	$70.27_{\pm 0.3}$	$77.55_{\pm 0.1}$	$80.14_{\pm 0.7}$	$57.60_{\pm 0.1}$	$66.14_{\pm 0.8}$	$69.04_{\pm 0.1}$	$23.41_{\pm 0.1}$	$30.44_{\pm 0.5}$	$32.60_{\pm 0.6}$
	Swin-B-1k	$71.20_{\pm 0.6}$	$78.74_{\pm 0.9}$	$80.38_{\pm 0.6}$	$58.66_{\pm 0.3}$	$68.43_{\pm 0.4}$	$71.29_{\pm 0.8}$	$25.30_{\pm 0.3}$	$34.50_{\pm 1.0}$	$34.81_{\pm 0.4}$
	Swin-B-22k	$70.99_{\pm 0.1}$	$78.85_{\pm 0.6}$	$80.19_{\pm 0.1}$	$56.12_{\pm 1.3}$	$68.31_{\pm 0.2}$	$70.14_{\pm 0.3}$	$25.26_{\pm 0.2}$	$33.88_{\pm 0.3}$	$34.88_{\pm 0.2}$
	Swin-L	$74.09_{\pm 3.0}$	$78.70_{\pm 0.9}$	$80.63_{\pm 0.4}$	$60.49_{\pm 1.7}$	$70.34_{\pm 1.9}$	$71.99_{\pm 1.6}$	$24.91_{\pm 0.2}$	$33.04_{\pm 0.7}$	$34.46_{\pm 0.3}$

4 Experiments

In this section, we employ state-of-the-art (SOTA) segmentation methods to examine the contributions and challenges of our newly curated RVD as well as establish a new benchmark for the dynamic vessel segmentation and localization tasks. As our data are collected from hand-held fundus imaging devices, we also investigate whether domain gaps between our dataset and existing ones which are captured by benchtop based devices.

4.1 Overall Results

We first focus on spatial segmentation tasks on our dataset. We conduct experiments of binary vessel segmentation, general artery-vein segmentation, and fine-grained artery-vein segmentation, respectively. We employ several popular used segmentation methods, including FCN [66], DeepLabV3 [11], Segmentor [73], and Mask2Former [12]. We also apply different backbones to these segmentation methods. The adopted backbones involves convolutional UNet [63], ResNet [30], ViT [20], and Swin Transformer [45]. We use the pre-trained parameters as initialization and train the networks on our training set. Due to the space limit, we only present the results of DeepLabV3 and Mask2Former in Table 2 and others are reported in the Appendix.

Even for the binary segmentation task, the highest mIoU results barely reach around 70%. For the more complex fine-grained artery-vein segmentation, the mIoU values further decline to approximately 25%. Some visual results are shown in Fig. 3. It is observed that current methods in general struggle to localize thin vessels. However, these thin vessels always play an important role in reflecting some diseases, *e.g.*, atherosclerosis [80]. The performance of the SOTA methods implies the challenges of our dataset, but this also emphasizes the potential of our dataset for future studies.

Table 3: Performance of SVP recognition and localization.

Method	Recognition			Localization		
	Acc	AUROC	Recall	F-mAP	V-mAP	mIOU
LRCN [19]	52.68	56.79	45.00	64.62	59.06	50.62
I3D [8]	60.71	61.83	61.67	67.49	60.63	51.89
X3D [24]	52.68	51.60	75.00	61.12	52.60	50.53
TSN [77]	50.89	64.39	30.00	67.60	56.85	50.89
VTN [55]	58.93	65.58	86.67	68.08	57.64	51.25

We also conduct experiments on temporal SVP recognition and localization with our provided annotations. In SVP recognition, we train the models to predict whether SVP exists in a video. In SVP localization, we train the models to identify the time period where SVP appears in a video. We employ the models of LRCN [19], I3D [8], X3D [24], TSN [77], and VTN [55]. We use the metrics in Section 3.4 and report the results in Table 3. To the best of our knowledge, we are the first to provide data and annotations for SVP recognition and localization. However, it is found that existing methods fail to recognize and localize SVP precisely on our real-clinic video data. For example, in SVP localization task, VTN only achieves 51.25% mIoU, which might not meet the needs of real-world applications. Such results indicate that more specific-designed methods are highly demanded.

4.2 Domain Gaps between RVD and Existing Datasets

We conduct a two-way evaluation process where models trained on our dataset are tested on previous datasets and models trained on existing datasets are also evaluated on our dataset. First, to evaluate

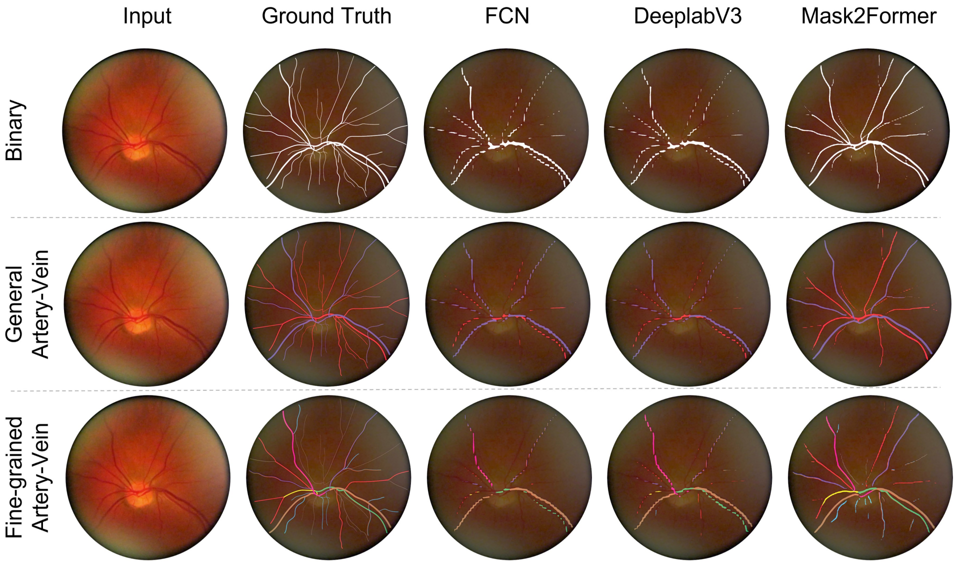


Figure 3: Visualization in the binary, general artery-vein, and fine-grained artery-vein segmentation.

Table 4: Evaluation of domain gaps between different datasets.

Method	Backbone	Binary Segmentation								General Artery-Vein	
		RVD ↓ C-DB.	C-DB. ↓ RVD	RVD ↓ DRI.	DRI. ↓ RVD	RVD ↓ HRF	HRF ↓ RVD	RVD ↓ FIVES	FIVES ↓ RVD	RVD ↓ RITE	RITE ↓ RVD
DLV3	UNet	$70.23_{\pm 0.2}$	$62.17_{\pm 0.4}$	$65.21_{\pm 0.2}$	$62.21_{\pm 0.3}$	$70.89_{\pm 0.3}$	$56.81_{\pm 0.5}$	$69.55_{\pm 0.2}$	$68.92_{\pm 0.4}$	$51.05_{\pm 0.2}$	$28.31_{\pm 0.4}$
	ResNet50	$71.56_{\pm 0.3}$	$63.41_{\pm 0.2}$	$66.34_{\pm 0.1}$	$63.04_{\pm 0.2}$	$71.27_{\pm 0.2}$	$57.76_{\pm 0.4}$	$68.73_{\pm 0.3}$	$64.95_{\pm 0.2}$	$51.84_{\pm 0.3}$	$29.87_{\pm 0.3}$
	ResNet101	$71.89_{\pm 0.4}$	$63.59_{\pm 0.2}$	$66.72_{\pm 0.3}$	$63.98_{\pm 0.1}$	$71.98_{\pm 0.3}$	$58.17_{\pm 0.9}$	$64.81_{\pm 0.2}$	$65.26_{\pm 0.3}$	$52.56_{\pm 0.2}$	$30.30_{\pm 0.4}$
M2F	ResNet50	$72.23_{\pm 0.2}$	$65.71_{\pm 0.3}$	$67.03_{\pm 0.2}$	$66.74_{\pm 0.1}$	$73.05_{\pm 0.2}$	$65.65_{\pm 0.2}$	$71.00_{\pm 1.5}$	$69.89_{\pm 0.4}$	$54.65_{\pm 0.2}$	$49.84_{\pm 0.3}$
	ResNet101	$73.57_{\pm 0.3}$	$67.20_{\pm 0.1}$	$67.57_{\pm 0.2}$	$67.25_{\pm 0.3}$	$73.76_{\pm 0.3}$	$65.66_{\pm 0.1}$	$65.27_{\pm 0.2}$	$69.39_{\pm 0.3}$	$55.23_{\pm 0.3}$	$50.12_{\pm 0.4}$
	Swin-T	$74.97_{\pm 0.2}$	$67.59_{\pm 0.3}$	$69.99_{\pm 0.3}$	$67.74_{\pm 0.3}$	$74.71_{\pm 0.1}$	$63.89_{\pm 0.3}$	$67.40_{\pm 0.3}$	$69.64_{\pm 0.1}$	$55.65_{\pm 0.4}$	$50.88_{\pm 0.2}$
	Swin-S	$75.21_{\pm 0.9}$	$67.98_{\pm 0.2}$	$69.93_{\pm 0.1}$	$67.51_{\pm 0.3}$	$73.97_{\pm 0.2}$	$67.05_{\pm 0.3}$	$72.87_{\pm 0.6}$	$69.53_{\pm 0.8}$	$56.47_{\pm 0.2}$	$51.42_{\pm 0.3}$
	Swin-B-1k	$74.93_{\pm 0.2}$	$68.13_{\pm 0.1}$	$71.21_{\pm 0.2}$	$67.64_{\pm 0.3}$	$76.03_{\pm 0.3}$	$66.88_{\pm 0.4}$	$71.30_{\pm 0.2}$	$69.57_{\pm 0.7}$	$57.04_{\pm 0.4}$	$52.99_{\pm 0.4}$
	Swin-B-22k	$76.95_{\pm 0.3}$	$70.04_{\pm 0.2}$	$73.79_{\pm 0.3}$	$68.36_{\pm 0.3}$	$78.67_{\pm 0.2}$	$64.06_{\pm 0.4}$	$73.74_{\pm 0.3}$	$69.34_{\pm 0.4}$	$57.05_{\pm 0.2}$	$52.84_{\pm 0.5}$
	Swin-L	$76.86_{\pm 1.3}$	$70.47_{\pm 0.2}$	$73.02_{\pm 0.1}$	$67.72_{\pm 0.4}$	$76.15_{\pm 0.3}$	$67.11_{\pm 0.3}$	$73.56_{\pm 0.9}$	$69.64_{\pm 1.5}$	$57.28_{\pm 0.4}$	$52.98_{\pm 0.3}$

binary vessel segmentation performance, we include the following datasets: CHASE DB1 (C-DB.), DRIVE (DRI.), HRF, and STARE (STA.). Then, we also conduct general artery-vein segmentation on the RITE dataset. The results are shown in Table 4. Note that existing datasets do not support fine-grained eight-class segmentation, and thus we did not test our data in this setting. Due to the domain gap, the models suffer performance drop. The results also indicate that our dataset provides unique data samples. The visualization is illustrated in the Appendix. From our experimental results above, we can tell that the retinal vessel segmentation is far from being solved. Our RVD dataset will serve as a valuable resource, motivating future explorations in retinal vessel segmentation.

5 Conclusion

In this work, we propose the first video-based retinal vessel segmentation dataset by employing handheld devices for data acquisition. Our dataset significantly complements the current benchtop-based datasets for retinal vessel segmentation and enables SVP detection and localization. More importantly, it offers rich annotations for both spatial vessel segmentation and temporal SVP localization. In comparison to existing datasets, our dataset is not only the largest scale one with the most diverse annotations but also more challenging. The domain gaps between our dataset and existing ones allows researchers to investigate how to minimize the domain gaps in vessel segmentation. Therefore, our curated dataset RVD is valuable for retinal vessel segmentation and would facilitate the clinical diagnosis of eye-related diseases.

Acknowledgement

This research is funded in part by ARC-Discovery grant (DP220100800) and ARC-DECRA grant (DE230100477). We thank all anonymous reviewers and ACs for their constructive suggestions.

References

[1] Samaneh Abbasi-Sureshjani, Iris Smit-Ockeloen, Jiong Zhang, and Bart Ter Haar Romeny. Biologically-inspired supervised vasculature segmentation in slo retinal fundus images. In *Image Analysis and Recognition: 12th International Conference, ICIAR 2015, Niagara Falls, ON, Canada, July 22-24, 2015, Proceedings 12*, pages 325–334. Springer, 2015.

[2] Michael D Abràmoff, Mona K Garvin, and Milan Sonka. Retinal imaging and image analysis. *IEEE reviews in biomedical engineering*, 3:169–208, 2010.

[3] Aziah Ali, W Mimi Diyana W Zaki, Aini Hussain, and Wan Haslina Wan Abdul Halim. Retinal width estimation of high-resolution fundus images for diabetic retinopathy detection. In *2020 IEEE-EMBS Conference on Biomedical Engineering and Sciences (IECBES)*, pages 460–465, 2021.

[4] Md Zahangir Alom, Mahmudul Hasan, Chris Yakopcic, Tarek M Taha, and Vijayan K Asari. Recurrent residual convolutional neural network based on u-net (r2u-net) for medical image segmentation. *arXiv preprint arXiv:1802.06955*, 2018.

[5] Ibrahim Atli and Osman Serdar Gedik. Sine-net: A fully convolutional deep learning architecture for retinal blood vessel segmentation. *Engineering Science and Technology, an International Journal*, 24(2):271–283, 2021.

[6] Buket D Barkana, Inci Saricicek, and Burak Yildirim. Performance analysis of descriptive statistical features in retinal vessel segmentation via fuzzy logic, ann, svm, and classifier fusion. *Knowledge-Based Systems*, 118:165–176, 2017.

[7] Roberto Brunelli. *Template matching techniques in computer vision: theory and practice*. John Wiley & Sons, 2009.

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

[9] Renoh Johnson Chalakkal, Waleed Habib Abdulla, and Sheng Chiong Hong. 3 - fundus retinal image analyses for screening and diagnosing diabetic retinopathy, macular edema, and glaucoma disorders. In Ayman S. El-Baz and Jasjit S. Suri, editors, *Diabetes and Fundus OCT*, Computer-Assisted Diagnosis, pages 59–111. Elsevier, 2020.

[10] Karen KW Chan, Fangyao Tang, Clement CY Tham, Alvin L Young, and Carol Y Cheung. Retinal vasculature in glaucoma: a review. *BMJ open ophthalmology*, 1(1):e000032, 2017.

[11] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[12] Bowen Cheng, Anwesa Choudhuri, Ishan Misra, Alexander Kirillov, Rohit Girdhar, and Alexander G Schwing. Mask2former for video instance segmentation. *arXiv preprint arXiv:2112.10764*, 2021.

[13] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022.

[14] Carol Yim-lui Cheung, Charumathi Sabanayagam, Antony Kwan-pui Law, Neelam Kumari, Daniel Shu-wei Ting, Gavin Tan, Paul Mitchell, Ching Yu Cheng, and Tien Yin Wong. Retinal vascular geometry and 6 year incidence and progression of diabetic retinopathy. *Diabetologia*, 60:1770–1781, 2017.

[15] Avijit Dasgupta and Sonam Singh. A fully convolutional neural network based structured prediction approach towards the retinal vessel segmentation. In *2017 IEEE 14th international symposium on biomedical imaging (ISBI 2017)*, pages 248–251. IEEE, 2017.

[16] Jyotiprava Dash and Nilamani Bhoi. A thresholding based technique to extract retinal blood vessels from fundus images. *Future Computing and Informatics Journal*, 2(2):103–109, 2017.

[17] Sonali Dash and Manas Ranjan Senapati. Enhancing detection of retinal blood vessels by combined approach of dwt, tyler coye and gamma correction. *Biomedical Signal Processing and Control*, 57:101740, 2020.

[18] Behdad Dashtbozorg, Ana Maria Mendonça, and Aurélio Campilho. An automatic graph-based approach for artery/vein classification in retinal images. *IEEE Transactions on Image Processing*, 23(3):1073–1083, 2013.

[19] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2625–2634, 2015.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[21] Rolando Estrada, Michael J Allingham, Priyatham S Mettu, Scott W Cousins, Carlo Tomasi, and Sina Farsiu. Retinal artery-vein classification via topology estimation. *IEEE transactions on medical imaging*, 34(12):2518–2534, 2015.

[22] Zhun Fan and Jia-Jie Mo. Automated blood vessel segmentation based on de-noising auto-encoder and neural network. In *2016 International conference on machine learning and cybernetics (ICMLC)*, volume 2, pages 849–856. IEEE, 2016.

[23] Damian JJ Farnell, Fraser N Hatfield, Paul Knox, Michael Reakes, Stan Spencer, David Parry, and Simon P Harding. Enhancement of blood vessels in digital fundus photographs via the application of multiscale line operators. *Journal of the Franklin institute*, 345(7):748–765, 2008.

[24] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020.

[25] Muhammad Moazam Fraz, Paolo Remagnino, Andreas Hoppe, Bunyarit Uyyanonvara, Alicja R Rudnicka, Christopher G Owen, and Sarah A Barman. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Transactions on Biomedical Engineering*, 59(9):2538–2548, 2012.

[26] Manuel E Gegundez-Arias, Diego Marin-Santos, Isaac Perez-Borrero, and Manuel J Vasallo-Vazquez. A new deep learning method for blood vessel segmentation in retinal images based on convolutional kernels and modified u-net model. *Computer Methods and Programs in Biomedicine*, 205:106081, 2021.

[27] Georgia Gkioxari and Jitendra Malik. Finding action tubes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 759–768, 2015.

[28] Mahdi Hashemzadeh and Baharak Adlpour Azar. Retinal blood vessel extraction employing effective image features and combination of supervised and unsupervised machine learning methods. *Artificial intelligence in medicine*, 95:1–15, 2019.

[29] Ali Hatamizadeh, Hamid Hosseini, Niraj Patel, Jinseo Choi, Cameron C Pole, Cory M Hoeferlin, Steven D Schwartz, and Demetri Terzopoulos. Ravir: A dataset and methodology for the semantic segmentation and quantitative analysis of retinal arteries and veins in infrared reflectance imaging. *IEEE Journal of Biomedical and Health Informatics*, 26(7):3272–3283, 2022.

[30] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[31] Qiao Hu, Michael D Abràmoff, and Mona K Garvin. Automated separation of binary overlapping trees in low-contrast color retinal images. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part II 16*, pages 436–443. Springer, 2013.

[32] Usama Iqbal. Smartphone fundus photography: a narrative review. *International Journal of Retina and Vitreous*, 7(1):44, 2021.

[33] Zhexin Jiang, Hao Zhang, Yi Wang, and Seok-Bum Ko. Retinal blood vessel segmentation using fully convolutional network with transfer learning. *Computerized Medical Imaging and Graphics*, 68:1–15, 2018.

[34] Kai Jin, Xingru Huang, Jingxing Zhou, Yunxiang Li, Yan Yan, Yibao Sun, Qianni Zhang, Yaqi Wang, and Juan Ye. Fives: A fundus image dataset for artificial intelligence based vessel segmentation. *Scientific Data*, 9(1):475, 2022.

[35] Aya F Khalaf, Inas A Yassine, and Ahmed S Fahmy. Convolutional neural networks for deep feature learning in retinal vessel segmentation. In *2016 IEEE international conference on image processing (ICIP)*, pages 385–388. IEEE, 2016.

[36] Tariq M Khan, Mohammad AU Khan, Naveed Ur Rehman, Khuram Naveed, Imran Uddin Afridi, Syed Saud Naqvi, and Imran Raazak. Width-wise vessel bifurcation for improved retinal vessel segmentation. *Biomedical Signal Processing and Control*, 71:103169, 2022.

[37] Ronald Klein, Chiu-Fang Chou, Barbara EK Klein, Xinzhi Zhang, Stacy M Meuer, and Jinan B Saaddine. Prevalence of age-related macular degeneration in the us population. *Archives of ophthalmology*, 129(1):75–80, 2011.

[38] Thomas Köhler, Attila Budai, Martin F Kraus, Jan Odstrčilik, Georg Michelson, and Joachim Hornegger. Automatic no-reference quality assessment for retinal fundus images using vessel segmentation. In *Proceedings of the 26th IEEE international symposium on computer-based medical systems*, pages 95–100. IEEE, 2013.

[39] KP Sampath Kumar, Debjit Bhowmik, G Harish, S Duraivel, and B Pragathi Kumar. Diabetic retinopathy-symptoms, causes, risk factors and treatment. *The Pharma Innovation*, 1(8), 2012.

[40] Avisek Lahiri, Vineet Jain, Arnab Mondal, and Prabir Kumar Biswas. Retinal vessel segmentation under extreme low annotation: A gan based semi-supervised approach. In *2020 IEEE international conference on image processing (ICIP)*, pages 418–422. IEEE, 2020.

[41] Di Li, Dhimas Arief Dharmawan, Boon Poh Ng, and Susanto Rahardja. Residual u-net for retinal vessel segmentation. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 1425–1429. IEEE, 2019.

[42] Wei Li, Mingquan Zhang, and Dali Chen. Fundus retinal blood vessel segmentation based on active learning. In *2020 International conference on computer information and big data applications (CIBDA)*, pages 264–268. IEEE, 2020.

[43] Paweł Liskowski and Krzysztof Krawiec. Segmenting retinal blood vessels with deep neural networks. *IEEE transactions on medical imaging*, 35(11):2369–2380, 2016.

[44] Bo Liu, Lin Gu, and Feng Lu. Unsupervised ensemble strategy for retinal vessel segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 111–119. Springer, 2019.

[45] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021.

[46] Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3431–3440, 2015.

[47] Wenao Ma, Shuang Yu, Kai Ma, Jiexiang Wang, Xinghao Ding, and Yefeng Zheng. Multi-task neural networks with spatial activation for retinal vessel segmentation and artery/vein classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part I 22*, pages 769–778. Springer, 2019.

[48] Yuhui Ma, Huaying Hao, Jianyang Xie, Huazhu Fu, Jiong Zhang, Jianlong Yang, Zhen Wang, Jiang Liu, Yalin Zheng, and Yitian Zhao. Rose: A retinal oct-angiography vessel segmentation dataset and new model. *IEEE Transactions on Medical Imaging*, 40(3):928–939, 2021.

[49] Debapriya Maji, Anirban Santara, Pabitra Mitra, and Debdoot Sheet. Ensemble of deep convolutional neural networks for learning to detect retinal vessels in fundus images. *arXiv preprint arXiv:1603.04833*, 2016.

[50] Ana Maria Mendonça, António Sousa, Luís Mendonça, and Aurélio Campilho. Automatic localization of the optic disc by combining vascular and intensity information. *Computerized medical imaging and graphics*, 37(5-6):409–417, 2013.

[51] Muthu Rama Krishnan Mookiah, Stephen Hogg, Tom J MacGillivray, Vijayaraghavan Prathiba, Rajendra Pradeepa, Viswanathan Mohan, Ranjit Mohan Anjana, Alexander S Doney, Colin NA Palmer, and Emanuele Trucco. A review of machine learning methods for retinal blood vessel segmentation and artery/vein classification. *Medical Image Analysis*, 68:101905, 2021.

[52] David Moreno-Ajona, James Alexander McHugh, and Jan Hoffmann. An update on imaging in idiopathic intracranial hypertension. *Frontiers in Neurology*, 11:453, 2020.

[53] William H Morgan, Martin L Hazelton, and Dao-Yi Yu. Retinal venous pulsation: Expanding our understanding and use of this enigmatic phenomenon. *Progress in retinal and eye research*, 55:82–107, 2016.

[54] Muhammad Muzammal Naseer, Kanchana Ranasinghe, Salman H Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Intriguing properties of vision transformers. *Advances in Neural Information Processing Systems*, 34:23296–23308, 2021.

[55] Daniel Neimark, Omri Bar, Maya Zohar, and Dotan Asselmann. Video transformer network. *arXiv preprint arXiv:2102.00719*, 2021.

[56] Meindert Niemeijer, Joes Staal, Bram Van Ginneken, Marco Loog, and Michael D Abramoff. Comparative study of retinal vessel segmentation methods on a new publicly available database. In *Medical imaging 2004: image processing*, volume 5370, pages 648–656. SPIE, 2004.

[57] Meindert Niemeijer, Xiayu Xu, Alina V Dumitrescu, Priya Gupta, Bram Van Ginneken, James C Folk, and Michael D Abramoff. Automated measurement of the arteriolar-to-venular width ratio in digital color fundus photographs. *IEEE Transactions on medical imaging*, 30(11):1941–1950, 2011.

[58] Américo Oliveira, Sergio Pereira, and Carlos A Silva. Retinal vessel segmentation based on fully convolutional neural networks. *Expert Systems with Applications*, 112:229–242, 2018.

[59] Pavle Prentašić, Sven Lončarić, Zoran Vatavuk, Goran Benčić, Marko Subašić, Tomislav Petković, Lana Dujmović, Maja Malenica-Ravlić, Nikolina Budimlija, and Rašeljka Tadić. Diabetic retinopathy image database (dridb): a new database for diabetic retinopathy screening programs research. In *2013 8th International Symposium on Image and Signal Processing and Analysis (ISPA)*, pages 711–716. IEEE, 2013.

[60] Touseef Ahmad Qureshi, Maged Habib, Andrew Hunter, and Bashir Al-Diri. A manually-labeled, artery/vein classified benchmark for the drive dataset. In *Proceedings of the 26th IEEE international symposium on computer-based medical systems*, pages 485–488. IEEE, 2013.

[61] Oscar Ramos-Soto, Erick Rodríguez-Esparza, Sandra E Balderas-Mata, Diego Oliva, Aboul Ella Hassanien, Ratheesh K Meleppat, and Robert J Zawadzki. An efficient retinal blood vessel segmentation in eye fundus images by using optimized top-hat and homomorphic filtering. *Computer Methods and Programs in Biomedicine*, 201:105949, 2021.

[62] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

[63] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

[64] Vallikutti Sathananthavathi and G Indumathi. Encoder enhanced atrous (eea) unet architecture for retinal blood vessel segmentation. *Cognitive Systems Research*, 67:84–95, 2021.

[65] Caroline A Schneider, Wayne S Rasband, and Kevin W Eliceiri. Nih image to imagej: 25 years of image analysis. *Nature methods*, 9(7):671–675, 2012.

[66] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):640–651, 2017.

[67] Hongwei Sheng, Xin Yu, Feiyu Wang, MD Khan, Hexuan Weng, Sahar Shariflou, and S Mojtaba Golzan. Autonomous stabilization of retinal videos for streamlining assessment of spontaneous venous pulsations. *arXiv preprint arXiv:2305.06043*, 2023.

[68] Seung Yeon Shin, Soochahn Lee, Il Dong Yun, and Kyoung Mu Lee. Deep vessel segmentation by learning graphical connectivity. *Medical image analysis*, 58:101556, 2019.

[69] João VB Soares, Jorge JG Leandro, Roberto M Cesar, Herbert F Jelinek, and Michael J Cree. Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification. *IEEE Transactions on medical Imaging*, 25(9):1214–1222, 2006.

[70] Jaemin Son, Sang Jun Park, and Kyu-Hwan Jung. Towards accurate segmentation of retinal vessels and the optic disc in fundoscopic images with generative adversarial networks. *Journal of digital imaging*, 32(3):499–512, 2019.

[71] Toufique Ahmed Soomro, Ahmed J Afifi, Junbin Gao, Olaf Hellwich, Lihong Zheng, and Manoranjan Paul. Strided fully convolutional neural network for boosting the sensitivity of retinal blood vessels segmentation. *Expert Systems with Applications*, 134:36–52, 2019.

[72] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE transactions on medical imaging*, 23(4):501–509, 2004.

[73] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

[74] Olubunmi Sule and Serestina Viriri. Enhanced convolutional neural networks for segmentation of retinal blood vessel image. In *2020 Conference on Information Communications Technology and Society (ICTAS)*, pages 1–6. IEEE, 2020.

[75] Peng Tang, Qiaokang Liang, Xintong Yan, Dan Zhang, Gianmarc Coppola, and Wei Sun. Multi-proportion channel ensemble model for retinal vessel segmentation. *Computers in biology and medicine*, 111:103352, 2019.

[76] Beaudelaire Saha Tchinda, Daniel Tchiotsop, Michel Noubom, Valerie Louis-Dorr, and Didier Wolf. Retinal blood vessels segmentation using classical edge detection filters and the neural network. *Informatics in Medicine Unlocked*, 23:100521, 2021.

[77] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[78] Jiahong Wei, Guijie Zhu, Zhun Fan, Jinchao Liu, Yibiao Rong, Jiajie Mo, Wenji Li, and Xinjian Chen. Genetic u-net: automatically designed deep networks for retinal vessel segmentation using a genetic algorithm. *IEEE Transactions on Medical Imaging*, 41(2):292–307, 2021.

[79] Maximilian WM Wintergerst, Divyansh K Mishra, Laura Hartmann, Payal Shah, Vinaya K Konana, Pradeep Sagar, Moritz Berger, Kaushik Murali, Frank G Holz, Mahesh P Shanmugam, et al. Diabetic retinopathy screening using smartphone-based fundus imaging in india. *Ophthalmology*, 127(11):1529–1538, 2020.

[80] Tien Yin Wong, FM Amirul Islam, Ronald Klein, Barbara EK Klein, Mary Frances Cotch, Cecilia Castro, A Richey Sharrett, and Eyal Shahar. Retinal vascular caliber, cardiovascular risk factors, and inflammation: the multi-ethnic study of atherosclerosis (mesa). *Investigative ophthalmology & visual science*, 47(6):2341–2350, 2006.

[81] Aaron Wu, Ziyue Xu, Mingchen Gao, Mario Buty, and Daniel J Mollura. Deep vessel tracking: A generalized probabilistic approach via deep learning. In *2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI)*, pages 1363–1367. IEEE, 2016.

[82] Rui Xu, Jiaxin Zhao, Xinchen Ye, Pengcheng Wu, Zhihui Wang, Haojie Li, and Yen-Wei Chen. Local-region and cross-dataset contrastive learning for retinal vessel segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part II*, pages 571–581. Springer, 2022.

[83] Dao-Yi Yu, K Yu Paula, Stephen J Cringle, Min H Kang, and Er-Ning Su. Functional and morphological characteristics of the retinal and choroidal vasculature. *Progress in Retinal and Eye Research*, 40:53–93, 2014.

[84] Jiong Zhang, Behdad Dashtbozorg, Erik Bekkers, Josien PW Pluim, Remco Duits, and Bart M ter Haar Romeny. Robust retinal vessel segmentation via locally adaptive derivative frames in orientation scores. *IEEE transactions on medical imaging*, 35(12):2631–2644, 2016.

[85] Mo Zhang, Fei Yu, Jie Zhao, Li Zhang, and Quanzheng Li. Befd: Boundary enhancement and feature denoising for vessel segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part V 23*, pages 775–785. Springer, 2020.

[86] Yishuo Zhang and Albert CS Chung. Deep supervision with additional labels for retinal vessel segmentation task. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2018: 21st International Conference, Granada, Spain, September 16-20, 2018, Proceedings, Part II 11*, pages 83–91. Springer, 2018.

[87] Yitian Zhao, Jianyang Xie, Huaizhong Zhang, Yalin Zheng, Yifan Zhao, Hong Qi, Yangchun Zhao, Pan Su, Jiang Liu, and Yonghuai Liu. Retinal vascular network topology reconstruction and artery/vein classification via dominant set clustering. *IEEE transactions on medical imaging*, 39(2):341–356, 2019.

[88] Chao Zhou, Xiaogang Zhang, and Hua Chen. A new robust method for blood vessel segmentation in retinal fundus images based on weighted line detector and hidden markov model. *Computer methods and programs in biomedicine*, 187:105231, 2020.

[89] Lei Zhou, Qi Yu, Xun Xu, Yun Gu, and Jie Yang. Improving dense conditional random field for retinal vessel segmentation by discriminative feature learning and thin-vessel enhancement. *Computer methods and programs in biomedicine*, 148:13–25, 2017.

APPENDIX

A Broader Impacts Statement

The creation and introduction of our video-based retinal vessel dataset (RVD) have profound effects on both the research community and the broader healthcare domain.

Research Impacts. By providing a comprehensive dataset with both spatial and temporal dimensions, the RVD significantly facilitates the analysis of retinal vessel segmentation and leads to improved understanding and modeling of ocular diseases. By incorporating dynamic video data, this dataset offers a broader and richer scope of retinal information than the traditional static image-based datasets. It fosters new research opportunities in retinal vessel segmentation that emphasize dynamic temporal characteristics and more granular vessel details. The introduced domain gaps with handheld devices may promote the development of robust and adaptable models, thereby advancing state-of-the-art image analysis methods. By providing this dataset as a resource for the research community, we hope to facilitate widespread collaboration and accelerate the exploration in retinal disease detection, understanding, and diagnosis.

Societal Impacts. As adopted in our RVD dataset, smartphone-based devices have the potential to democratize retinal vessel examination by making it more accessible and less reliant on expensive, specialized ophthalmic equipment. Such an enhancement in accessibility consequently leads to earlier detection and prevention of a spectrum of ocular diseases. Handheld devices potentially increase equity in healthcare services.

Limitations. Although our RVD is the largest dataset for retinal vessel segmentation to date (635 videos with annotations), its scale is still limited compared to other datasets in computer vision and thus our dataset can be further extended in the future. Compared to the data collected with bench-top devices, the original videos captured with handheld devices involve more realistic factors such as operator techniques, varying lighting conditions, and eye movement of the patient. These factors will require more sophisticated data cleaning and preprocessing strategies to avoid the degraded quality and reliability of the data.

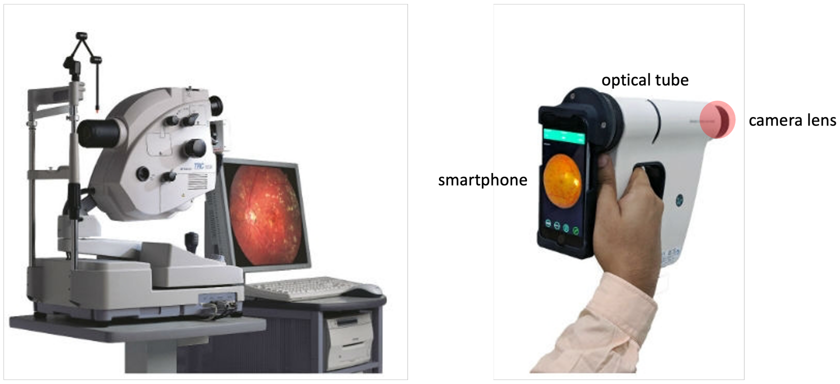
B Building RVD

Equipment. We use self-designed handheld devices for data collection. Compared to bench-top devices which are cumbersome and expensive, the devices we adopt here are lightweight and portable. Our devices are much cheaper and easier to access. Our devices are built by connecting a smartphone to the fundus camera lens via an optical tube (see Fig. 1).

Operation details. Clinicians use the handheld devices in Fig. 1 (b) to amass a collection of videos during eye health examinations. To accommodate handheld operation, each video lasts at least 0.5 seconds and does not exceed 25 seconds. The process initiates with the random selection of participants and their consent prior to recording. Then, either the left eye, the right eye, or both eyes are randomly selected for video recording. This method ensures that our dataset comprises both healthy and diseased eyes. Each participant joins in the recording process either once or multiple times. In our dataset, we try to eliminate the potential bias towards specific eye conditions and ensure a broad representation of various ocular health states from different people.

Statistics of our RVD. We show some statistical characteristics of the patients in Fig. 3. Age distribution among patients exhibits a Gaussian-like one, and more videos are collected from males rather than females in our dataset. From Fig. 3 (c), we conclude that most of the videos are collected from clinic 0 and clinic 3. Finally, the videos collected from the left eye and right eye are nearly balanced, as shown in Fig. 3 (d). More detailed statistical information on the videos can be accessed on our website in the future.

Characteristics of our videos. We show more samples in Fig. 2 to illustrate the diversity of our collected videos, *e.g.*, samples with different sizes of optic disc regions (ODR), different illumination, and different vessel density. The ODR contains most of the vessels in the retina and the size of ODR is vital in retinal vessel segmentation. Different illumination and vessel density will also make our dataset more challenging by increasing the variety of samples.



(a) Bench-top device

(b) An example of our handheld devices

Figure 1: (**a**): Bench-top device, which is cumbersome and expensive; (**b**): Our handheld device, which is lightweight and portable. It is built by connecting a smartphone to the fundus camera lens via a optical tube.

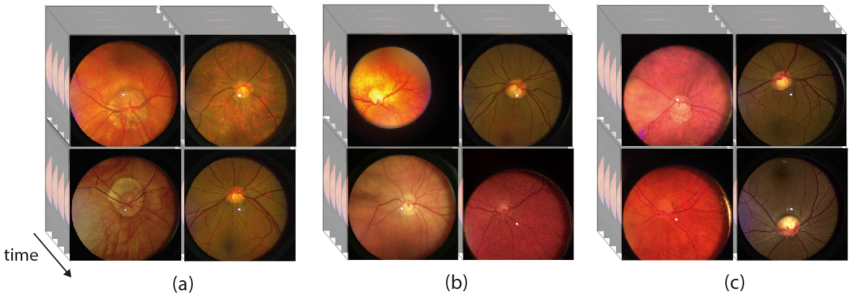


Figure 2: Diversity of our collected videos. (**a**): Samples with different sizes of ODR; (**b**): Different illumination; (**c**): Different vessel density.

Statistics of diseased and normal eyes. As indicated in Table 1, there are 317 videos of diseased eyes and 318 videos of normal eyes. Although the number of videos collected from different clinics varies, *e.g.*, only dozens of videos are collected from clinics Q and R, while more than 200 videos are collected from clinics P and S, the diseased and normal eye videos for each clinic are balanced.

Table 1: The statistics of diseased and normal eyes.

Clinics	P	Q	R	S	Total
Diseased	132	18	16	151	317
Normal	140	15	42	121	318

C Experimental settings

C.1 Implementation Details.

We implement the benchmark in PyTorch using the open-sourced MMSegmentation [1]. For all methods, we leverage the default settings of each method in MMSegmentation and implement them on 4090 GPUs. We train each model for 40,000 iterations and select the checkpoints with the best validation results. To ensure the accuracy and reliability of final results, cross-validation is employed across our experiments.

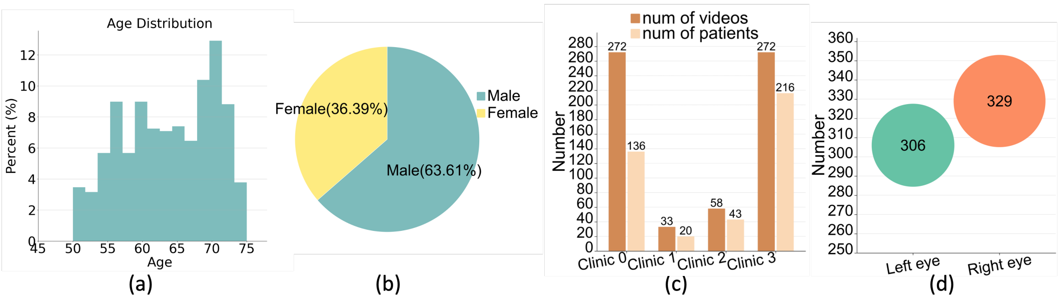


Figure 3: Statistics of our dataset. **(a)**: The age distribution of the participants; **(b)**: The ratio of males and females; **(c)**: The number of videos and patients in each clinic; **(d)**: The number of videos collected from the left eye and right eye.

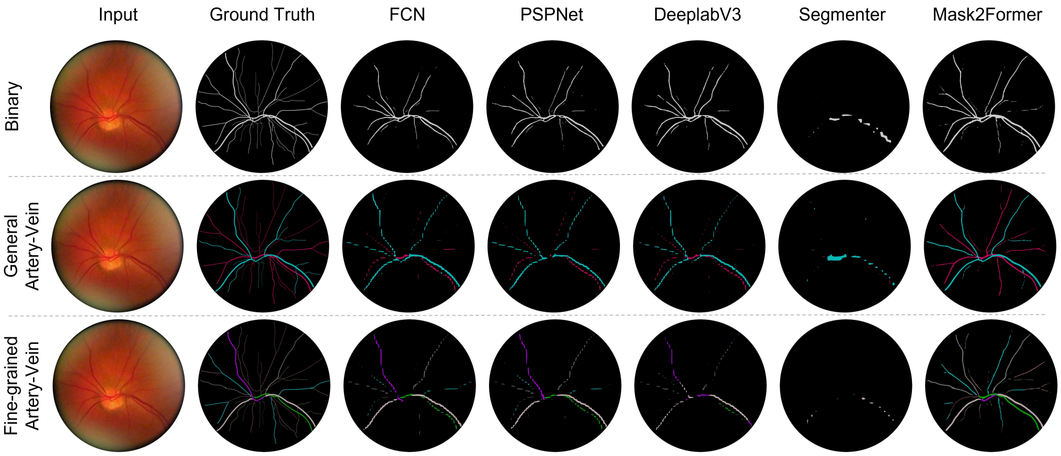


Figure 4: Visualization in the binary, general artery-vein, and fine-grained artery-vein segmentation.

C.2 More Results of Models

In Table 2, we present the segmentation results produced by FCN, PSPNet, and Segmenter. Each method leverages different backbones, *e.g.*, UNet, ResNet, and ViT. The models are trained and tested with binary vessel masks, general artery-vein masks, and fine-grained artery-vein masks. We observe a consistent pattern that even though the methods have achieved the best results with binary masks, the highest mIoU is under 70. These results underscore the difficulties of existing methods in dealing with our dataset. Such results show the challenges inherent to our dataset and imply the potential of our dataset to inspire future studies. In Fig. 4, we show the complete segmentation results of FCN, PSPNet, DeeplabV3, Segmenter, and Mask2Former in our dataset.

C.3 Better domain gap analysis

In this section, we investigate the potential domain gaps in our RVD dataset. Such variations in distribution could arise from multiple factors: **(i)** Different clinical conditions; **(ii)** Variations of vasculature due to the presence and absence of disease and **(iii)** Different devices used for video capture. Furthermore, we consider the evaluation of baseline models on a third dataset.

C.3.1 Test across disease and normal videos

We evaluate our best-performing model Mask2Former with the Swin Transformer backbone on diseased and normal retinal videos. The binary segmentation results are reported in Table 3. The segmentation results on both diseased and normal eye videos are close, indicating the domain discrepancy between the diseased and normal videos is marginal in the context of retinal vessel segmentation.

Table 2: Segmentation results of “FCN”, “PSPNet”, and “Segmenter” on our RVD dataset.

Method	Backbone	Binary			General Artery-Vein			Fine-grained Artery-Vein		
		mIoU	mAcc	mFscore	mIoU	mAcc	mFscore	mIoU	mAcc	mFscore
FCN [5]	UNet [4]	$67.82_{\pm0.6}$	$73.22_{\pm0.6}$	$77.08_{\pm0.5}$	$38.29_{\pm1.0}$	$39.85_{\pm0.9}$	$64.59_{\pm0.7}$	$13.47_{\pm0.5}$	$14.88_{\pm0.9}$	$19.95_{\pm1.2}$
	ResNet50 [3]	$62.12_{\pm0.5}$	$66.05_{\pm0.7}$	$70.76_{\pm0.3}$	$49.22_{\pm0.1}$	$53.93_{\pm0.2}$	$58.99_{\pm0.1}$	$18.38_{\pm0.6}$	$21.41_{\pm0.0.8}$	$26.62_{\pm0.4}$
	ResNet101	$62.79_{\pm0.3}$	$66.77_{\pm0.4}$	$71.54_{\pm0.3}$	$48.24_{\pm0.6}$	$51.98_{\pm0.5}$	$57.90_{\pm0.5}$	$18.53_{\pm0.3}$	$21.44_{\pm0.1}$	$24.07_{\pm0.2}$
PSPNet [7]	UNet	$68.53_{\pm0.5}$	$74.04_{\pm0.6}$	$77.80_{\pm0.1}$	$40.08_{\pm0.8}$	$42.27_{\pm0.2}$	$45.65_{\pm0.1}$	$12.71_{\pm0.7}$	$13.67_{\pm0.6}$	$64.34_{\pm0.4}$
	ResNet50	$61.82_{\pm0.3}$	$65.25_{\pm0.2}$	$70.37_{\pm0.5}$	$49.08_{\pm0.6}$	$53.94_{\pm0.9}$	$58.71_{\pm1.1}$	$18.92_{\pm0.8}$	$22.10_{\pm1.2}$	$24.45_{\pm1.1}$
	ResNet101	$63.06_{\pm0.6}$	$67.11_{\pm0.2}$	$71.87_{\pm0.1}$	$47.76_{\pm0.3}$	$51.34_{\pm0.8}$	$57.12_{\pm0.5}$	$19.37_{\pm1.3}$	$22.39_{\pm2.0}$	$25.05_{\pm1.5}$
Segmenter [6]	ViT-T [2]	$49.39_{\pm1.3}$	$51.25_{\pm1.2}$	$51.51_{\pm0.9}$	$33.83_{\pm0.1}$	$35.15_{\pm0.3}$	$36.04_{\pm0.5}$	$11.98_{\pm0.2}$	$12.57_{\pm0.1}$	$29.73_{\pm0.4}$
	ViT-S	$51.36_{\pm0.4}$	$53.33_{\pm0.7}$	$55.14_{\pm0.1}$	$32.54_{\pm0.5}$	$33.79_{\pm1.0}$	$33.61_{\pm0.2}$	$11.62_{\pm0.5}$	$12.11_{\pm0.4}$	$28.37_{\pm0.9}$
	ViT-B	$50.98_{\pm0.3}$	$52.90_{\pm2.3}$	$54.45_{\pm0.7}$	$34.03_{\pm0.4}$	$35.36_{\pm1.3}$	$36.40_{\pm0.9}$	$11.78_{\pm1.1}$	$12.30_{\pm0.4}$	$28.99_{\pm1.2}$
	ViT-L	$48.11_{\pm0.3}$	$50.00_{\pm0.5}$	$98.07_{\pm1.1}$	$34.70_{\pm1.3}$	$36.03_{\pm0.8}$	$37.55_{\pm0.7}$	$12.19_{\pm0.1}$	$12.75_{\pm0.3}$	$24.37_{\pm0.6}$

Table 3: Binary segmentation results of Mask2Former on diseased/normal videos.

Eye Condition	Diseased	Normal
mIoU	74.25	73.96
mAcc	79.23	78.16
mFscore	81.18	80.08

C.3.2 Test across different devices

In our dataset, videos are acquired by using two types of camera models. To study the domain gaps between the data collected from different devices, we focus on the binary segmentation task and adopt the best-performing model Mask2Former with the Swin transformer backbone. Specifically, we train the model on data collected from one device and test it on that of another device. The results are shown in Table 4. The performance of the models varies slightly across different devices. This suggests that the variations in the data collection process introduce some domain gaps.

Table 4: Results of models trained on one device and tested on another device.

Device	Metric	Device 1	Device 2
Device 1	mIoU	69.37	71.07
	mAcc	77.07	80.45
	mFscore	78.73	80.16
Device 2	mIoU	68.53	70.98
	mAcc	74.2	77.99
	mFscore	77.85	80.04

C.3.3 Test across different clinics

To study model performance across different clinics, we focus on binary segmentation and SVP detection. Here, we denote the four clinics as P, Q, R, and S. We adopt the same model in Section C.3.1 and test it on each clinic test data. The results are reported in Table 5 and Table 6. In both vessel segmentation and SVP detection, the model performances across different clinics are different, indicating that domain gaps exist in different clinics.

Table 5: Binary segmentation results of Mask2Former across different clinics.

Clinics	P	Q	R	S
mIoU	73.85	72.00	70.89	75.33
mAcc	79.63	77.72	75.63	81.78
mFscore	81.35	80.54	78.38	82.27

C.3.4 Test on the third dataset

We further conduct experiments to evaluate the domain gaps between our RVD dataset and existing datasets. To be specific, we train segmentation models on RVD and DRIVE with the same amount of samples and then test them on a third dataset CHASEDB. As seen in Table 7, models trained on RVD have lower performance compared to those trained on DRIVE. This indicates our dataset exhibits a larger domain gap with the existing datasets.

Table 6: Evaluation of the I3D model on SVP detection across various clinics.

Clinics	P	Q	R	S
Acc	64.58	50.00	60.00	58.33
AUROC	70.28	66.67	58.83	54.32
Recall	65.38	66.67	90.00	51.85

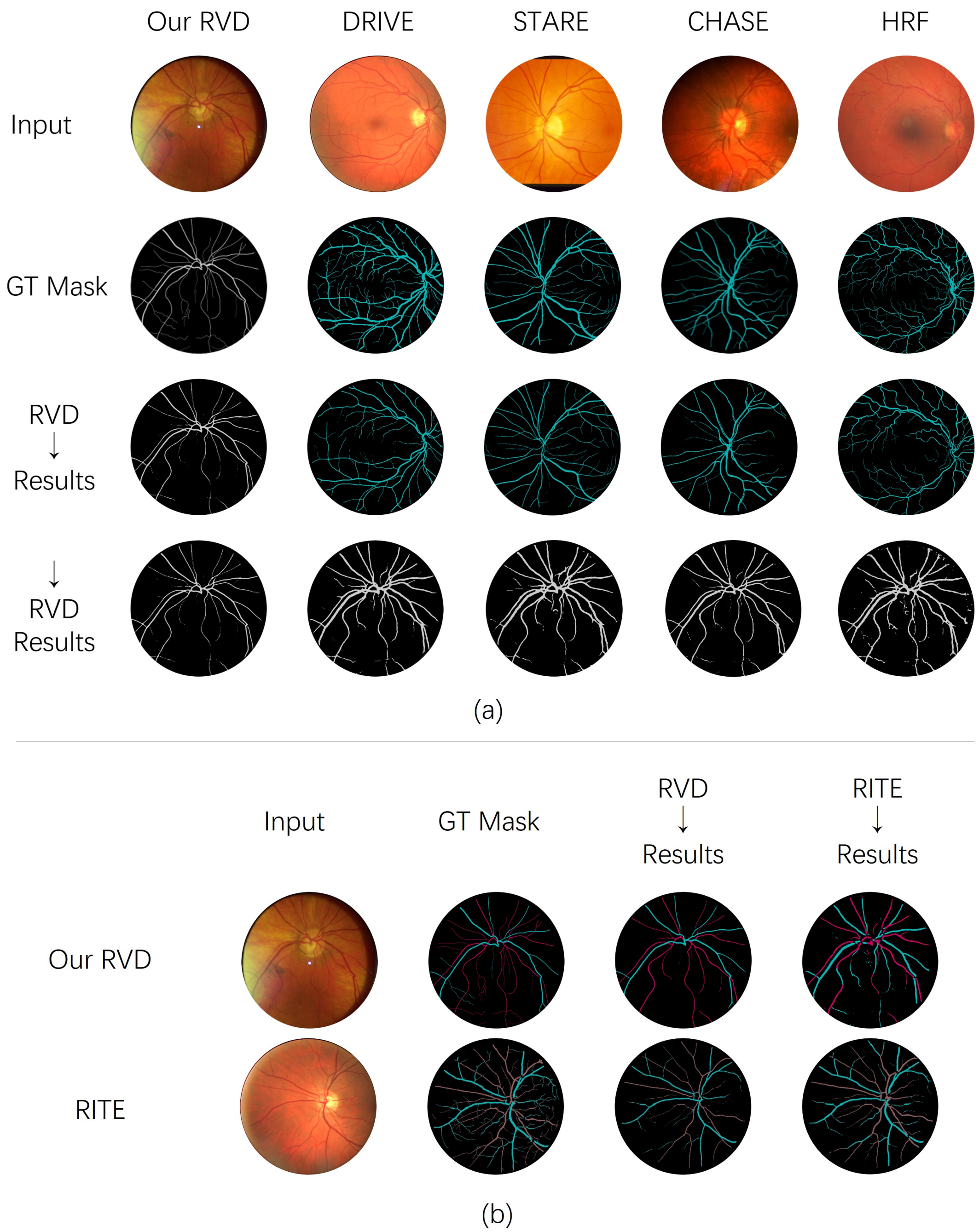


Figure 5: Visualization of domain gaps between different datasets. **(a)**: Examples of binary segmentation datasets; **(b)**: Visualization results of general Artery-Vein segmentation datasets.

C.4 Domain Gaps between RVD and Existing Datasets

In Section 4.2, we highlight the presence of domain gaps between existing datasets and our retinal vessel dataset (RVD). To further demonstrate this phenomenon, we show a set of visualization results, which can be found in Fig. 5. We consider two distinct scenarios:

(a) Initially, we present the visualization results of models trained on existing datasets and then applied to our dataset. The results reveal a concerning trend of presence of overgeneralization in the predictions, thereby overlooking finer details. This underscores the difficulty for models trained on existing datasets to generalize to our dataset.

Table 7: Evaluation of domain gaps between RVD and existing datasets. The models are trained on RVD and DRIVE and tested on CHASEDB.

Model-Backbone	RVD → CHASEDB			DRIVE → CHASEDB		
	mIoU	mAcc	mFscore	mIoU	mAcc	mFscore
DeepLabV3-R50	60.22	63.3	68.25	65.56	76.00	75.69
Mask2Former-R50	68.35	76.00	77.62	78.48	89.43	86.84
Mask2Former-Swin-L	63.47	67.60	73.04	79.88	90.97	87.87

(b) Conversely, we also show the visualization of models initially trained on our RVD, and then applied to existing datasets. Similar to the first case, the performance drop is observed after transferring. However, the transferred results reveal that more granular details, particularly of vessel structure, are preserved. Such a phenomenon suggests that models trained on our dataset exhibit much better generalizability and tend to adapt more efficiently to the existing datasets.

C.5 Computational Efficiency

We analyze the computational operations (i.e., GFLOPs) and inference time (i.e., ms) for trained models. Specifically, for models trained with spatial annotations, we analyze Mask2Former and DeepLabv3 with different backbones. The results are presented as follows in Table 8. The overall speed of different models is faster for processing our data. Models with larger backbones require more computational cost and more inference time, for example, from Swin-T to Swin-L, both GFLOPs and the inference time are increased a lot. However, we notice that the inference time may not linearly grow with GFLOPs across different models, since the inference time can be affected by many factors, *e.g.*, CUDA optimization for different network structures.

For models trained with temporal annotations, we consider SVP detection with different backbones. We provide their GFLOPs and inference time in Table 9. The speed of models for SVP detection is still faster, *e.g.*, LRCN processes each video in 89 ms. However, when compared with models in binary segmentation, their speed can be further improved. These results indicate that more sophisticated model structures are needed for our datasets.

Table 8: Computational efficiency of Mask2Former and DeepLabv3 with different backbones.

Model	Mask2Former							DeepLabv3	
Backbone	ResNet50	ResNet101	Swin-T	Swin-S	Swin-B-1k	Swin-B-22k	Swin-L	ResNet50	ResNet101
GFLOPs (G)	40	59	28	36	86	86	160	270	347
Inference time (ms)	27.30	29.7	31.3	45.3	67.2	62.8	92.2	24.5	33.7

Table 9: Computational efficiency of models trained with temporal annotations.

Model	LRCN	I3D	X3D	TSN	VTN
GFLOPs	16	263	180	252	1374
Inference time (ms)	89.0	226.7	182.8	322.3	486.2

References

[1] MMSegmentation Contributors. MMSegmentation: Openmmlab semantic segmentation toolbox and benchmark. `https://github.com/open-mmlab/mmsegmentation`, 2020.

[2] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[4] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th*

International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, pages 234–241. Springer, 2015.

[5] Evan Shelhamer, Jonathan Long, and Trevor Darrell. Fully convolutional networks for semantic segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 39(4):640–651, 2017.

[6] Robin Strudel, Ricardo Garcia, Ivan Laptev, and Cordelia Schmid. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7262–7272, 2021.

[7] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.