APPENDIX

Paper ID: 223

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.

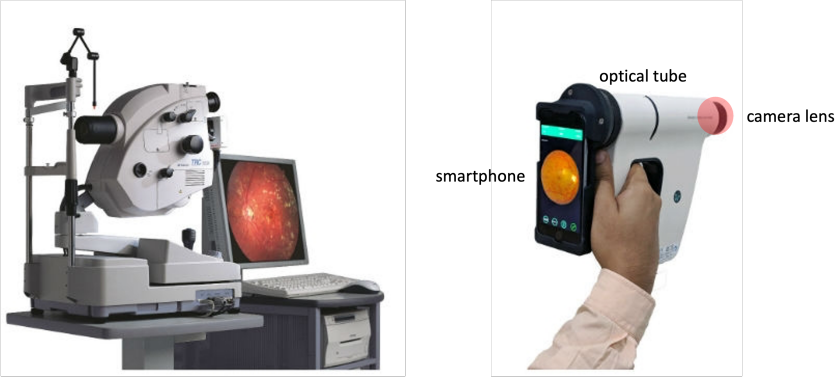
Limitations. Although our RVD is the largest dataset for retinal vessel segmentation to date (635 videos with annotations), its scale is still limited compared to other datasets in computer vision and thus our dataset can be further extended in the future. Compared to the data collected with bench-top 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.



(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.

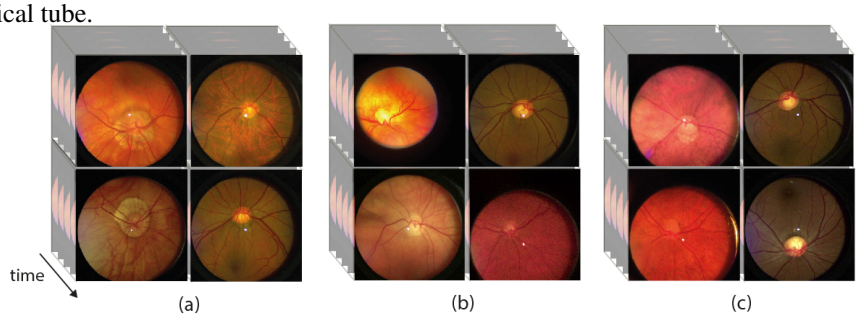


Figure 2: Diversity of our collected videos. **(a)**: Samples with different sizes of ODR; **(b)**: Different illumination; **(c)**: Different vessel density.

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.

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

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.

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.

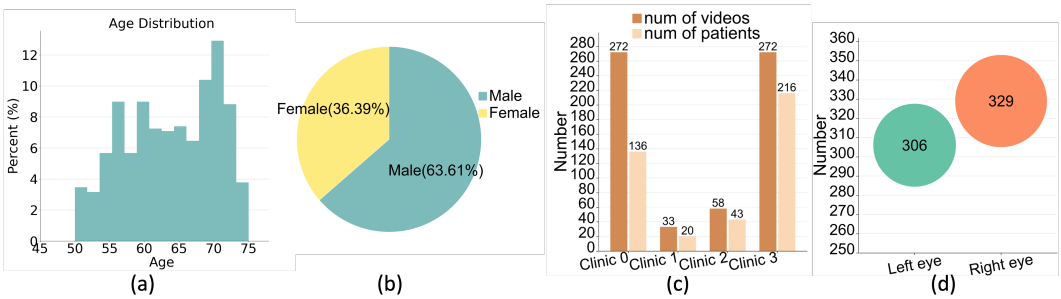


Figure 3: Statistics of our dataset. **(a)**: The age distribution of the participants; **(b)**: The ratio of males and females; **(c)**: The number of videos and patients in each clinic; **(d)**: The number of videos collected from the left eye and right eye.

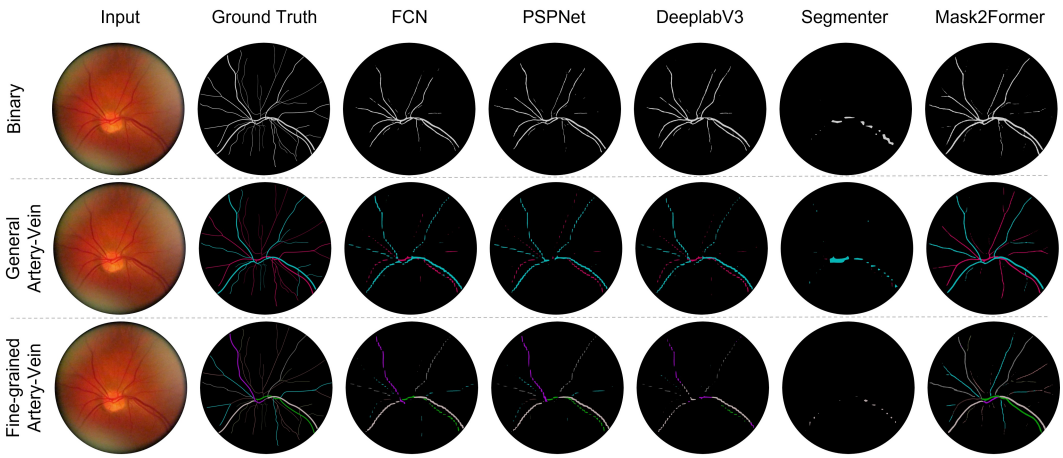


Figure 4: Visualization in the binary, general artery-vein, and fine-grained artery-vein segmentation.

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 seg-

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]	$67.82_{\pm 0.6}$	$73.22_{\pm 0.6}$	$77.08_{\pm 0.5}$	$38.29_{\pm 1.0}$	$39.85_{\pm 0.9}$	$64.59_{\pm 0.7}$	$13.47_{\pm 0.5}$	$14.88_{\pm 0.9}$	$19.95_{\pm 1.2}$
	ResNet50 [3]	$62.12_{\pm 0.5}$	$66.05_{\pm 0.7}$	$70.76_{\pm 0.3}$	$49.22_{\pm 0.1}$	$53.93_{\pm 0.2}$	$58.99_{\pm 0.1}$	$18.38_{\pm 0.6}$	$21.41_{\pm 0.0.8}$	$26.62_{\pm 0.4}$
	ResNet101	$62.79_{\pm 0.3}$	$66.77_{\pm 0.4}$	$71.54_{\pm 0.3}$	$48.24_{\pm 0.6}$	$51.98_{\pm 0.5}$	$57.90_{\pm 0.5}$	$18.53_{\pm 0.3}$	$21.44_{\pm 0.1}$	$24.07_{\pm 0.2}$
PSPNet [7]	UNet	$68.53_{\pm 0.5}$	$74.04_{\pm 0.6}$	$77.80_{\pm 0.1}$	$40.08_{\pm 0.8}$	$42.27_{\pm 0.2}$	$45.65_{\pm 0.1}$	$12.71_{\pm 0.7}$	$13.67_{\pm 0.6}$	$64.34_{\pm 0.4}$
	ResNet50	$61.82_{\pm 0.3}$	$65.25_{\pm 0.2}$	$70.37_{\pm 0.5}$	$49.08_{\pm 0.6}$	$53.94_{\pm 0.9}$	$58.71_{\pm 1.1}$	$18.92_{\pm 0.8}$	$22.10_{\pm 1.2}$	$24.45_{\pm 1.1}$
	ResNet101	$63.06_{\pm 0.6}$	$67.11_{\pm 0.2}$	$71.87_{\pm 0.1}$	$47.76_{\pm 0.3}$	$51.34_{\pm 0.8}$	$57.12_{\pm 0.5}$	$19.37_{\pm 1.3}$	$22.39_{\pm 2.0}$	$25.05_{\pm 1.5}$
Segmenter [6]	ViT-T [2]	$49.39_{\pm 1.3}$	$51.25_{\pm 1.2}$	$51.51_{\pm 0.9}$	$33.83_{\pm 0.1}$	$35.15_{\pm 0.3}$	$36.04_{\pm 0.5}$	$11.98_{\pm 0.2}$	$12.57_{\pm 0.1}$	$29.73_{\pm 0.4}$
	ViT-S	$51.36_{\pm 0.4}$	$53.33_{\pm 0.7}$	$55.14_{\pm 0.1}$	$32.54_{\pm 0.5}$	$33.79_{\pm 1.0}$	$33.61_{\pm 0.2}$	$11.62_{\pm 0.5}$	$12.11_{\pm 0.4}$	$28.37_{\pm 0.9}$
	ViT-B	$50.98_{\pm 0.3}$	$52.90_{\pm 2.3}$	$54.45_{\pm 0.7}$	$34.03_{\pm 0.4}$	$35.36_{\pm 1.3}$	$36.40_{\pm 0.9}$	$11.78_{\pm 1.1}$	$12.30_{\pm 0.4}$	$28.99_{\pm 1.2}$
	ViT-L	$48.11_{\pm 0.3}$	$50.00_{\pm 0.5}$	$98.07_{\pm 1.1}$	$34.70_{\pm 1.3}$	$36.03_{\pm 0.8}$	$37.55_{\pm 0.7}$	$12.19_{\pm 0.1}$	$12.75_{\pm 0.3}$	$24.37_{\pm 0.6}$

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

mentation 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.

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

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

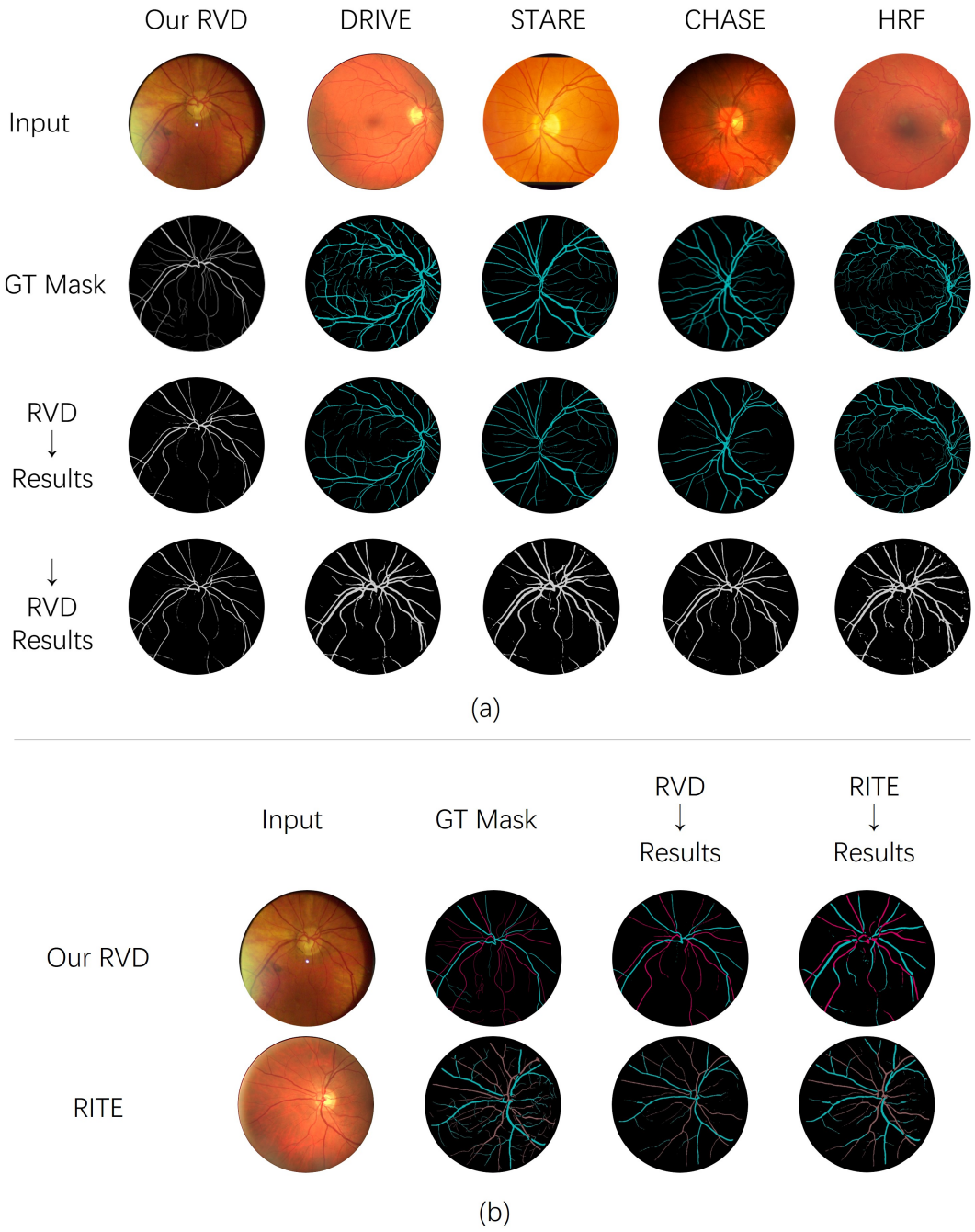


Figure 5: Visualization of domain gaps between different datasets. **(a)**: Examples of binary segmentation datasets; **(b)**: Visualization results of general Artery-Vein segmentation datasets.

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.

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:

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

(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.

(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.

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.