BioMedInformatics

MDPI

Article

Toward Modality-Aware Preprocessing for Deep Learning-Based Blood Vessel Segmentation

Sinan AlKassar [1,*] and Haval Tariq Sadeeq [2]

[1] Research and Development Directorate, Ministry of Higher Education and Scientific Research, Baghdad 10001, Iraq
[2] Artificial Intelligence Department, Technical College of Duhok, Duhok Polytechnic University, Duhok 42001, Kurdistan Region, Iraq; haval.tariq@dpu.edu.krd
* Correspondence: sinan.alkassar@mohesr.edu.iq

Abstract

Background: Preprocessing pipelines developed for retinal fundus photography are routinely transferred to other vascular imaging modalities without validation. **Methods:** Seven preprocessing stages from the retinal literature were assembled into a rule-based chain gated by image-quality descriptors calibrated from training-set percentiles. A ten-condition ablation (the complete chain, seven chains each omitting one stage, and two minimal chains of two stages each) with five-fold cross-validation was run on VessMAP, 100 confocal images of mouse cortical vasculature, and six conditions were repeated on the DRIVE and CHASE_DB1 fundus datasets. **Results:** On VessMAP, only removing intensity normalisation significantly reduced Dice (-0.017, paired Wilcoxon $p < 10^{-4}$); removing local contrast normalisation raised it by 0.019, and two-stage chains exceeded the complete chain by 0.025–0.027. Removing local contrast normalisation or denoising, or reducing the chain to two stages, also improved Dice on both fundus datasets. With thresholds fitted on training folds only, VessMAP Dice was 0.882 ± 0.006, within the inter-annotator interval [0.864, 0.891], and a 59,034-parameter network came within 0.012 Dice of a 20.9-million-parameter model. **Conclusions:** The complete chain was not the best configuration on any dataset, including the fundus datasets its stages came from; preprocessing should be validated stage by stage on the target data.

Keywords: blood vessel segmentation; ablation study; image preprocessing; domain generalisation; centreline Dice

1. Introduction

Quantitative analysis of cortical vascular networks underpins the study of cerebral blood flow, neurovascular coupling and the vascular contribution to neurodegenerative disease. Properties such as vessel diameter, branch-point density, tortuosity and network connectivity are all derived from a binary vessel mask, so segmentation errors propagate directly into the downstream physiological measurements [1,2]. High-resolution confocal fluorescence microscopy of mouse brain vasculature now produces such masks routinely, and the accuracy of the segmentation step, therefore, determines the reliability of the morphometry built upon it.

Automated vessel segmentation has been studied most extensively on retinal fundus photography, where a small number of public benchmarks has supported two decades of methodological development. Early work classified pixels using hand-crafted descriptors,

Check for updates

Academic Editor: James C. L. Chow

Received: 7 September 2026
Revised: 7 October 2026
Accepted: 15 October 2026
Published: 19 October 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

including multi-scale Gabor features [3] and ensembles of bagged and boosted decision trees over richer feature sets [4], alongside filter-based methods such as the B-COSFIRE family [5,6] and matched filtering with Gaussian derivatives [7]. The introduction of U-Net [8] shifted the field to end-to-end convolutional segmentation, and subsequent architectures refined the decoder through nested dense skip connections [9], channel and spatial attention [10,11] and hard attention on difficult pixels [12]. More recently, transformer and state-space hybrids have been proposed to model global context [13–16], and topology-aware losses and architectures have been introduced to preserve vessel connectivity [17–21].

A consistent observation across this literature is that architectural capacity is not the binding constraint on the small datasets typical of vessel segmentation. Minimalistic networks with tens of thousands of parameters match architectures an order of magnitude larger [22–24], which suggests that the data, the loss and the preprocessing determine mask quality more than encoder depth does. Preprocessing in particular has settled into a stable set of conventions. Contrast limited adaptive histogram equalisation [25], Frangi vesselness filtering [26], wavelet denoising [27], bilateral filtering [28], homomorphic illumination correction [29] and local contrast normalisation appear across published pipelines, frequently with parameters fixed by the author rather than calibrated to the dataset [3,4,7,30,31]. Where these parameters have been questioned, the response has generally been to optimise them [32] rather than to test whether the operation is warranted at all.

These conventions are increasingly exported beyond the fundus domain. Work on domain adaptation has transferred retinal models to new datasets through pseudo-labelling [33] and diffusion-based appearance translation [34], and work on domain generalisation has done so through meta-learning on anatomy-consistent pseudo-modalities [35], shape-aware representations [36], domain-oriented feature embedding [37] and structural-stylistic augmentation [38], in several cases across imaging modalities. Cortical vessel segmentation, by contrast, has been studied predominantly on three-dimensional volumes from light-sheet microscopy [39] and angiography [40], and the release of VessMAP [41] has only recently made two-dimensional cortical confocal fluorescence microscopy accessible as a benchmark.

Cortical confocal fluorescence microscopy differs from fundus photography in three respects that bear directly on preprocessing. First, vessel density varies widely: the vessel-density measure distributed with VessMAP (vessel skeleton length per unit image area) ranges from 0.012 to 0.078 across its 100 images, a more than sixfold range. Second, the images are sparse and dark, with mean intensity near 13% of the dynamic range and a background of structured noise at scales comparable to the vessel calibre. Third, and most consequentially, the polarity of vessel contrast is reversed: retinal vessels absorb green light and appear darker than the surrounding retina in the green channel used by most fundus pipelines, whereas fluorescently labelled cortical vessels appear brighter than the parenchyma.

Despite the routine transfer of retinal preprocessing pipelines to new vascular imaging modalities, evidence remains limited on which preprocessing stages remain beneficial after a modality change. Existing domain generalisation methods adapt the model while leaving the preprocessing chain that precedes it unexamined, and the modality pairs they consider do not isolate the effect of preprocessing when preprocessing assumptions change with the imaging domain. Whether an operation calibrated for fundus photography helps, does nothing, or actively harms segmentation on a modality that violates these assumptions, therefore, requires direct stage-wise evaluation rather than a single aggregate score for a complete pipeline.

We address this gap by assembling seven preprocessing stages drawn from the retinal literature into a chain whose gate thresholds are calibrated from percentile statistics of the training set, and by measuring stage contributions on cortical confocal fluorescence microscopy and, for a reduced condition set, on DRIVE and CHASE_DB1 fundus images. The chain is deliberately rule-based: its gates are fixed rules on hand-crafted quality descriptors, and it neither learns nor detects the imaging modality. "Modality-aware" in the title, therefore, describes the practice the study argues for, selecting preprocessing operations only after they have been validated on the target modality, rather than a property of the chain, which serves as the instrument with which transferability is measured. The contribution of the study, therefore, lies in the stage-wise analysis of which operations survive a change of dataset and modality, not in the chain as a new enhancement method. The principal contributions are as follows:

- A stage-wise ablation, with paired per-image tests, on one cortical confocal and two retinal fundus datasets, showing that a complete chain assembled from the retinal literature is not the best configuration on any of them: removing local contrast normalisation or denoising, or reducing the chain to two stages, improves Dice on the fundus datasets the stages were designed for as well as on the cortical one.
- A quantitative characterisation of vessel polarity as a descriptor examined alongside cross-modality transfer failure, using a threshold-free separability measure rather than overlap scores alone.
- Evidence on model-size efficiency obtained by comparing networks spanning approximately 2.55 orders of magnitude (about 354-fold) in parameter count under a fixed preprocessing chain and training protocol.
- An analysis of how the evaluation protocol can inflate reported segmentation accuracy, together with a threshold-selection procedure that reduces this source of inflation without retraining.

The remainder of this paper is organised as follows. Section 2 reviews related work on retinal vessel segmentation, preprocessing for vascular imaging, cortical vessel segmentation and domain generalisation. Section 3 describes the datasets, the polarity measurement, the quality features and the adaptive preprocessing chain, the segmentation model, the inference and threshold-selection protocol, and the ablation and cross-dataset designs. Section 4 reports the cross-validation results, the per-stage ablation, the multi-dataset ablation, the architectural comparison, and the cross-dataset generalisation experiments. Section 5 interprets these findings and states the limitations of the study, and Section 6 concludes.

2. Related Work

2.1. Retinal Vessel Segmentation

The retinal vessel segmentation literature has developed on a small set of public benchmarks—DRIVE [42], STARE [43], CHASE_DB1 [4], HRF [44] and more recently FIVES [45]—which share colour fundus photography as their modality. Supervised pixel-level classification dominated the early literature [3], elaborated with ensembles of bagged and boosted decision trees over hand-crafted features [4]. Filter-based methods of comparable accuracy include the B-COSFIRE family [5,6] and matched filtering with first-order Gaussian derivatives [7].

End-to-end convolutional segmentation began with U-Net [8]. DeepVessel combined a convolutional network with a conditional random field [46]; a fully convolutional network operating on whole images gave faster inference at similar accuracy [47]; and EfficientNet encoders entered the field early [30]. U-Net++ [9] introduced nested dense skip connections that let the decoder fuse encoder features at multiple depths, which suits thin capillar-

ies whose signal exists only at the highest spatial frequencies. Squeeze-and-excitation blocks [10] re-weight channels by global pooled responses, and the spatial-and-channel variant (scSE) [11] applies spatial re-weighting in parallel. Hard attention [12] routes difficult pixels to a separate decoder branch.

In terms of transformer and state-space architectures, CoVi-Net interleaves convolutional and vision transformer blocks [13]; a Swin-Transformer/Res2Net hybrid targets microvessel localisation [14]; Mamba variants combine frequency-selective attention with gated convolutions [15]. Lightweight hybrid transformers developed for other medical segmentation tasks show that compact architectures can match much larger ones when inference speed matters [16].

A recurring concern is that architectural complexity has diminishing returns on the small training sets typical of vessel segmentation. A two-level W-shaped network with tens of thousands of parameters matched architectures one to three orders of magnitude larger on DRIVE, CHASE_DB1 and HRF [22]; a systematic analysis of few-parameter networks reached the same conclusion [23]; and the parameter floor was pushed to 78,000 with a skeleton-distance loss [24]. Dual-encoder filtering architectures continue this line [48].

A second concern is that pixel-wise Dice does not penalise breaks in the predicted vessel network if the broken segments are small; centreline Dice (clDice) [17] applies skeletonisation to both prediction and ground truth before computing overlap, which makes it sensitive to exactly those breaks. Complementary work pursues topology preservation architecturally, through connectivity-mask branches [18], breakpoint attention [19], skeleton-guided supervision [20] and topology-aware re-weighting [21].

2.2. Preprocessing for Vascular Imaging

Preprocessing for retinal vessel segmentation has accumulated a stable set of operations; it is their transferability, however, that is in question. Contrast limited adaptive histogram equalisation (CLAHE) [25] redistributes pixel intensities within local tiles; most published retinal pipelines [3,4,7,30,31] apply it without ablation, with tile size and clip limit fixed by the author rather than calibrated to the dataset. Where the parameters have been questioned, the response has been to optimise rather than to test them. Contrast enhancement of fundus images by particle swarm optimisation in [32] notes explicitly that studies apply CLAHE with default contextual region and clip limit.

Frangi's multiscale vesselness filter [26] analyses eigenvalues of the local Hessian to produce a ridge-like response at tubular structures. It has a bright-ridge and a dark-ridge mode, and the choice between them is dictated by the modality. Frangi-based enhancement has been used as a structural prior in retinal segmentation networks [49], and improved Frangi variants remain an active line of work [50].

Wavelet soft-threshold denoising [27] suppresses high-frequency content below an estimated noise floor, and bilateral filtering [28] preserves edges while smoothing uniform regions. Both target the frequency band where thin capillary boundaries also live. Homomorphic illumination correction [29] separates surface signal from the illumination component in log space.

Local Contrast Normalisation (LCN) subtracts a locally estimated mean and divides by a locally estimated standard deviation. In fundus images, where illumination varies smoothly over distances larger than a vessel diameter, LCN suppresses the illumination gradient while preserving the vessel boundary. In confocal fluorescence microscopy of cortical capillaries, illumination is delivered through an objective and is comparatively uniform, so there is little gradient to suppress.

The present chain should be distinguished from two related lines of work. Adaptive and optimised enhancement methods tune the parameters of operations chosen in advance,

for example, by particle swarm optimisation of contrast enhancement [32] or by comparing preprocessing variants on retinal benchmarks [51], so that image quality or downstream accuracy is maximised within one modality. Self-configuring frameworks such as nnU-Net [52] set intensity normalisation and resampling from a dataset fingerprint by fixed rules. The chain studied here is also rule-based, but it addresses a different question: not how to parameterise an operation, but whether the operation should be applied at all once the dataset or modality changes, which is answered by removing it.

2.3. Cortical Vessel Segmentation

Before VessMAP, cortical vessel segmentation was studied predominantly on three-dimensional volumes, most notably whole-mouse-brain light-sheet acquisitions [39], with volumetric architectures reporting strong within-modality performance on centreline and bifurcation prediction [40]. Neither was evaluated on confocal images of the cortical surface, which have a different noise model, much sparser vessel distributions and substantially higher heterogeneity across samples. VessMAP is the first publicly available benchmark for cortical confocal fluorescence vessel segmentation, with a reported inter-annotator Dice of 0.878 ± 0.031.

2.4. Domain Generalisation in Vessel Segmentation

Cross-dataset generalisation has been studied primarily within the retinal fundus modality. Minimalistic networks evaluated under strict leave-one-dataset-out (LODO) conditions achieve Dice between 0.55 and 0.72 on held-out datasets without fine-tuning [22]. Benchmarking across FIVES found image quality variations to be a stronger driver of performance drops than patient disease state [53]. Domain adaptation methods use target-domain images during training, for example dual-teacher pseudo-labelling [33] and diffusion-based appearance translation [34]. Domain generalisation methods do not; they rely on meta-learning on anatomy-consistent pseudo-modalities [35], shape-aware representations [36], Hessian-based vector fields [54], domain-oriented feature embedding [37] or structural-stylistic augmentation [38]. Ensemble pseudo-labels have also been used to train a student network [55].

Learning theory bounds the error on a target domain by the source error plus a measure of divergence between the two domains [56], and the medical domain adaptation literature reviewed in [57] largely consists of ways to reduce that divergence in feature space. Preprocessing changes the input distribution before any of these mechanisms act, which is why its transferability is examined separately here. Several of these methods address cross-modality shift: augmentation-based domain generalisation has been validated from fundus training to optical coherence tomography angiography test sets, and anatomy-consistent meta-learning enumerates cross-modality shift explicitly [35]. Prior work has also compared preprocessing choices directly for retinal vessel segmentation [51], and evaluated contrast-enhancement, morphological, and wavelet preprocessing variants [58,59]. Other learned models have incorporated preprocessing-derived features into their input representation [60]. These studies motivate treating preprocessing as an experimental variable, whereas the present work examines its behaviour across a confocal cortical dataset and retinal fundus datasets with different contrast characteristics.

3. Materials and Methods

3.1. Datasets

VessMAP [41] provides 100 greyscale confocal fluorescence microscopy images of mouse brain vasculature at 256×256 pixels, sampled from a base pool of 18,279 patches using a coreset-style protocol designed to maximise heterogeneity along four quality

dimensions: contrast, noise, vessel density and medial-line heterogeneity. The dataset's vessel-density measure ranges from 0.012 to 0.078 (mean 0.042), and mean pixel intensity is approximately 13% of the dynamic range. Every image carries a single expert annotation; a subset of 20 images was annotated independently by a second expert.

DRIVE [42] contains 40 colour fundus photographs at 565×584 pixels, of which 20 were used here; STARE [43] contains 20 at 700×605 pixels; CHASE_DB1 [4] contains 28 at 999×960 pixels. In each case the first annotator's masks are used.

3.2. Vessel Polarity

Because several preprocessing operations and the interpretation of cross-modality transfer both depend on whether vessels are brighter or darker than their surroundings, polarity was measured rather than assumed. For each dataset, mean intensity inside the expert annotation was compared against mean intensity in a dilated ring around it, with out-of-field pixels excluded.

Excluding the field of view is essential. Fundus images are a circular retina on a black surround, and comparing vessels against all non-vessel pixels drags the background mean below the vessel mean and reports dark vessels as bright. On a synthetic fundus image with vessels at intensity 85 against a retina at 120, the naive comparison returns $+7.4$ (bright, incorrect) while the local-ring comparison returns -35.0 (dark, correct).

Table 1 reports the measurement. Values are mean intensity inside the expert annotation minus mean intensity in a dilated ring surrounding it, with out-of-field pixels excluded, averaged over 12 images per dataset. The naive column compares against all non-vessel pixels and is reported to show the field-of-view artefact it introduces. VessMAP vessels are bright and all three fundus datasets have dark vessels. The Frangi filter is, therefore, configured per dataset.

Table 1. Measured vessel polarity.

Dataset	Vessel − Local Background	Naive Comparison	Polarity
VessMAP	+18.72	+19.05	Bright
DRIVE	−8.82	+22.35	Dark
CHASE_DB1	−9.62	+14.41	Dark
STARE	−19.77	+2.46	Dark

3.3. Quality Feature Estimation

Four scalar descriptors are estimated from each image without access to its annotation. They drive the conditional gates of the preprocessing chain. Let $\hat{I} = I/255$ denote the unit-normalised image.

Root-mean-square contrast is the standard deviation of pixel intensities, $c = \mathrm{std}(\hat{I})$. High-frequency noise is the standard deviation of the difference between the image and its Gaussian blur, $\sigma_n = \mathrm{std}(\hat{I} - G_1 * \hat{I})$. Vessel density is the foreground fraction under Otsu thresholding, $\rho = \mathrm{mean}(\hat{I} > \tau)$ for bright-vessel modalities and $\mathrm{mean}(\hat{I} < \tau)$ for dark-vessel modalities, with polarity taken from Table 1. Spatial heterogeneity is the coefficient of variation of a difference-of-Gaussians image, $h = \mathrm{std}(\Delta)/(|\mathrm{mean}(\Delta)| + \varepsilon)$ with $\Delta = G_3 * \hat{I} - G_1 * \hat{I}$.

A fifth descriptor, illumination variation $v = \mathrm{std}(G_r * \hat{I})$ with $r = \lfloor \min(H, W)/8 \rfloor$, gates the homomorphic stage. Because $\mathrm{mean}(\Delta)$ approaches zero by construction, h is heavy-tailed and its absolute value is not comparable across datasets; gating is percentile-based and, therefore, depends only on its ranking.

3.4. Adaptive Preprocessing Chain

Figure 1 shows the seven preprocessing stages in execution order. Four are conditional, activated only when a quality feature crosses a dataset-specific threshold computed once from the percentile distribution of features across the training images,

$$\tau_f = \text{Percentile}\Big(\{f(q_i)\}_{i=1}^{N},\, p_f\Big), \tag{1}$$

so the operating points of the chain adapt to any image collection within a modality without manual tuning. Which stages to include, however, depends on the target data and is determined here by ablation rather than assumed.

Stage 1, intensity normalisation, runs on every image. Pixel values are clipped to the $[p_2, p_{98}]$ intensity range and linearly rescaled,

$$x \leftarrow \frac{\text{clip}(I/255,\, p_2,\, p_{98}) - p_2}{p_{98} - p_2 + \varepsilon}, \tag{2}$$

which removes extreme outliers caused by specular reflections, dust on the coverslip, or sensor hot pixels.

Stage 2, homomorphic illumination correction, activates when v exceeds the 75th percentile of illumination variation; the image is then high-pass filtered in log space with a Gaussian kernel of $\sigma = \max(10,\, 0.1 \min(H, W))$ pixels following the standard homomorphic recipe [29]. On VessMAP 25% of images activate this stage.

Stage 3, adaptive CLAHE, activates when contrast c falls below the 15th percentile, $\tau_c = 0.024$ on VessMAP. The clip limit of the CLAHE operator scales with the contrast deficit,

$$\gamma_c = 1.5 + 2.5 \cdot \frac{\tau_c - c}{\tau_c + \varepsilon}, \tag{3}$$

with an 8×8 tile grid. Around 15% of VessMAP images activate this stage.

Stage 4 cascades bilateral filtering (window $d = 9$, $\sigma_c = \sigma_s = 30$) with soft-threshold wavelet denoising, activated when high-frequency noise exceeds the 75th percentile ($\tau_n = 0.016$ on VessMAP, 25% of images). The bilateral parameters are the conventional OpenCV setting, with σ_c in 8-bit intensity units and σ_s in pixels; they were not tuned, because the ablation asks whether the stage is useful at all. Wavelet denoising uses the Haar (Daubechies-1) wavelet with BayesShrink soft thresholding, the noise level estimated from the finest diagonal sub-band, and three fewer decomposition levels than the maximum the image size allows.

Stage 5 blends the Frangi vesselness response into the working image,

$$x \leftarrow (1 - \alpha)\, x + \alpha\, V, \tag{4}$$

with the filter evaluated at scales $\sigma \in \{1, 2, 3\}$ pixels in the ridge mode matching the dataset polarity and normalised to $[0, 1]$, and $\alpha = 0.25$ in the complete chain. Section 4.2 reports a sweep over the blend weight α.

Stage 6 applies local contrast normalisation with a 15×15 uniform filter for the local mean μ_{15} and local standard deviation σ_{15},

$$x \leftarrow \frac{x - \mu_{15}}{\sigma_{15} + \varepsilon}, \tag{5}$$

rescaled to $[0, 1]$ afterwards, activated when heterogeneity h exceeds the 60th percentile (around 40% of images).

Stage 7 calibrates the image by a power law that targets a mean intensity of 0.5,

$$\gamma = \mathrm{clip}\left(\frac{\log 0.5}{\log(\bar{x}+\varepsilon)}, 0.4, 2.5\right), \qquad x \leftarrow x^{\gamma}, \tag{6}$$

with $\bar{x}$ the current image mean. Algorithms 1 and 2 list the feature estimation and the chain in execution order, and Table 2 collects every parameter together with the image property each stage assumes.

Algorithm 1 Per-image quality feature estimation.

Input: greyscale image I; dataset polarity flag BRIGHT
Output: feature vector $[c,\ \sigma_n,\ \rho,\ h,\ v]^{\top}$

$\hat{I} \leftarrow I/255$
$c \leftarrow \mathrm{std}(\hat{I})$
$\sigma_n \leftarrow \mathrm{std}(\hat{I} - G_1 * \hat{I})$
$\tau \leftarrow \mathrm{OtsuThreshold}(\hat{I})$
if BRIGHT **then**
    $\rho \leftarrow \mathrm{mean}(\hat{I} > \tau)$
else
    $\rho \leftarrow \mathrm{mean}(\hat{I} < \tau)$
end if
$\Delta \leftarrow G_3 * \hat{I} - G_1 * \hat{I}$
$h \leftarrow \mathrm{std}(\Delta)/(|\mathrm{mean}(\Delta)| + \varepsilon)$
$v \leftarrow \mathrm{std}\left(G_{\lfloor \min(H,W)/8 \rfloor} * \hat{I}\right)$
return $[c,\ \sigma_n,\ \rho,\ h,\ v]^{\top}$

Table 2. Parameters of the preprocessing chain, the condition under which each stage runs, and the image property it assumes. Shares of activated images refer to VessMAP; all gates use the image-only descriptors of Algorithm 1. H, W: image height and width.

Stage	Operation and Parameters	Gate (VessMAP Share)	Property Assumed	Setting That Changes with Modality
S1	Clip to $[p_2, p_{98}]$, linear rescale to $[0,1]$	Always	Intensity range, outlier pixels (acquisition)	None
S2	Homomorphic high-pass in log domain, Gaussian $\sigma = \max(10, 0.1\min(H,W))$	$v > p_{75}$ (25%)	Slowly varying illumination (acquisition)	Gate threshold
S3	CLAHE, 8×8 tiles, clip limit $1.5 + 2.5(\tau_c - c)/\tau_c$	$c < p_{15}$, $\tau_c = 0.024$ (15%)	Low global contrast (acquisition)	Gate threshold
S4	Bilateral ($d = 9$, $\sigma_c = \sigma_s = 30$), then Haar wavelet, BayesShrink soft threshold	$\sigma_n > p_{75}$, $\tau_n = 0.016$ (25%)	Noise finer than vessels (acquisition)	Gate threshold
S5	Frangi, $\sigma \in \{1,2,3\}$, normalised to $[0,1]$, blend $\alpha = 0.25$	Always	Tubular vessels of known polarity (structure)	Ridge polarity (Table 1)
S6	Local contrast normalisation, 15×15 window, rescale to $[0,1]$	$h > p_{60}$ (40%)	Local contrast varying over more than a vessel width (illumination)	Gate threshold
S7	Power law to mean 0.5, $\gamma \in [0.4, 2.5]$	Always	Global intensity distribution (acquisition)	None

The order follows the convention of the retinal pipelines from which the stages are drawn: corrections of acquisition properties (intensity range, illumination, global contrast and noise; stages 1–4) precede structural enhancement (stage 5), and the operations that reshape the intensity distribution passed to the network (stages 6–7) come last. Five of the seven stages target properties of image formation and are generic in form; only Frangi blending encodes vascular structure, and it is the one stage whose configuration must change with the modality, through the ridge polarity of Table 1. Local contrast normalisation

sits between the two groups: it corrects illumination, but over a fixed window whose size relative to the vessel calibre differs between modalities. Gate thresholds are recomputed for every dataset.

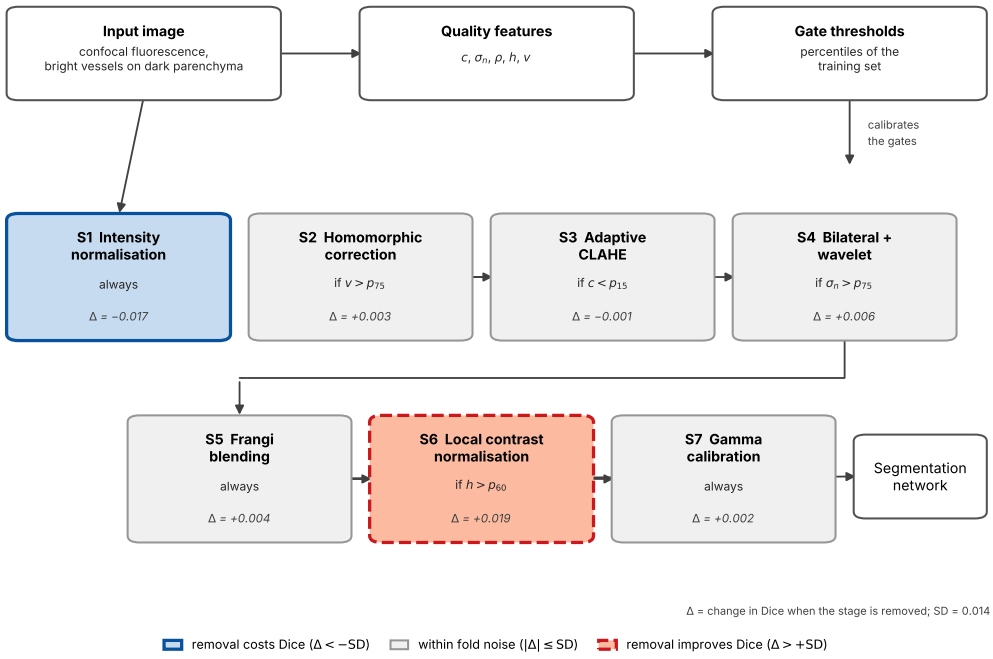


Figure 1. Block diagram of the input stage and the adaptive preprocessing chain. Shading encodes the five-fold ablation verdict of Section 4.2 on VessMAP: stages whose removal costs Dice, stages whose removal changes Dice by less than the inter-fold standard deviation, and stages whose removal improves Dice. White boxes show the input image, the estimation of quality features and gate thresholds, and the segmentation network; arrows give the order of processing, and the dashed red frame marks the stage whose removal improves Dice.

Algorithm 2 Adaptive preprocessing chain. Stages 1, 5 and 7 always run; stages 2, 3, 4 and 6 are gated on the quality features of the image.

Input: image I; features $[c, \sigma_n, \rho, h, v]$; thresholds $[\tau_c, \tau_n, \tau_h, \tau_v]$; polarity flag BRIGHT
Output: preprocessed image x

$x \leftarrow \mathrm{clip}(I/255, p_2, p_{98})$; $x \leftarrow (x - p_2)/(p_{98} - p_2 + \varepsilon)$ ▷ S1 intensity normalisation
if $v > \tau_v$ **then** ▷ S2 homomorphic correction
  $x \leftarrow \exp(\log(1 + 255x) - G_\sigma * \log(1 + 255x))$; rescale to $[0, 1]$
end if
if $c < \tau_c$ **then** ▷ S3 adaptive CLAHE
  $\gamma_c \leftarrow 1.5 + 2.5(\tau_c - c)/(\tau_c + \varepsilon)$; $x \leftarrow \mathrm{CLAHE}(x, \gamma_c, 8{\times}8)$
end if
if $\sigma_n > \tau_n$ **then** ▷ S4 denoising
  $x \leftarrow \mathrm{WaveletDenoise}(\mathrm{Bilateral}(x; d{=}9, \sigma_c{=}30, \sigma_s{=}30))$
end if
$V \leftarrow \mathrm{Frangi}(x,\ \sigma \in \{1, 2, 3\},\ \mathrm{black_ridges} = \neg\mathrm{BRIGHT})$
$x \leftarrow (1 - \alpha)x + \alpha V$ ▷ S5 Frangi blending
if $h > \tau_h$ **then** ▷ S6 local contrast normalisation
  $x \leftarrow (x - \mu_{15})/(\sigma_{15} + \varepsilon)$; rescale to $[0, 1]$
end if
$\gamma \leftarrow \mathrm{clip}(\log 0.5/\log(\bar{x} + \varepsilon),\ 0.4,\ 2.5)$; $x \leftarrow x^\gamma$ ▷ S7 gamma calibration
return x

3.5. Segmentation Model

The backbone is U-Net++ [9] with nested dense skip connections at four encoder depths and an ImageNet-pretrained EfficientNet [61] encoder. Section 4.4 compares EfficientNet-B0 (6.6 M parameters) against EfficientNet-B4 (20.9 M) under an otherwise identical protocol; the cross-validation results of Section 4.1 and the ablation use EfficientNet-B4, and B0 is recommended for further use because it performs equivalently. The decoder applies an scSE attention block [11] after each upsampling step. Before training, all batch normalisation layers in the encoder are replaced with instance normalisation layers with learnable affine parameters, which prevents the model from memorising running statistics specific to a single dataset.

The loss combines focal loss [62] with Tversky loss [63],

$$L = 0.4\, L_{\text{Focal}}(\gamma = 2, \varepsilon = 0.05) + 0.6\, L_{\text{Tversky}}(\alpha, \beta), \tag{7}$$

with label smoothing $\varepsilon = 0.05$ to prevent overconfident predictions at boundaries that are genuinely ambiguous between annotators [64]. The Tversky weights are derived from a dataset-level mean vessel density $\bar{\rho}$, the average over all 100 images of the per-image vessel-density measure distributed with VessMAP [41]; this is a different quantity from the Otsu foreground fraction ρ of Section 3.3,

$$\alpha = 0.15 + 2.5(\bar{\rho} - 0.04), \quad \beta = 1 - \alpha, \quad \bar{\rho} \in [0.04, 0.10]. \tag{8}$$

Equivalently, α and β interpolate linearly between $(0.15, 0.85)$ at $\bar{\rho} = 0.04$ and $(0.30, 0.70)$ at $\bar{\rho} = 0.10$, with $\bar{\rho}$ clipped to that range. The VessMAP value used in all reported experiments is $\bar{\rho} = 0.042$, which gives $\alpha = 0.155$ and $\beta = 0.845$, penalising missed vessels more heavily than spurious detections. The asymmetry reflects the priority for morphometry: missed capillaries break the network topology used by downstream measurements, whereas small false-positive fragments are easier to filter post hoc. Two auxiliary 1×1 convolutional heads are attached to the decoder at one-half and one-quarter of the output resolution following the deeply-supervised nets recipe [65]. Each auxiliary head contributes $0.3\times$ its own loss to the total,

$$L_{\text{total}} = L + 0.3 \sum_{s \in \{1/2,\, 1/4\}} L(\hat{y}_s,\, \downarrow_s y), \tag{9}$$

where $\downarrow_s y$ denotes the ground-truth mask downsampled to scale s by average pooling.

All models train with Adam [66] at weight decay 10^{-5} under a one-cycle schedule with peak learning rate 10^{-3}, 10% linear warm-up and cosine annealing, for 15 epochs per fold at batch size 4 with mixed precision and gradient norm clipped at 1.0; this schedule applies to all cross-validation experiments, and the cross-modality experiments use the 30-epoch schedule given in Section 3.9. Every cross-validation and LODO model is evaluated with its final-epoch weights, and no checkpoint is selected on test data; the cross-modality models of Section 3.9 are selected on a source-domain validation split. Augmentation applies random flips, rotation up to 45°, brightness and contrast jitter at ± 0.2, Gaussian noise at $\sigma \leq 0.05$ and coarse dropout. For LODO training, a Fourier-domain style mixing transform [67] replaces the low-frequency amplitude spectrum of a training image with a mixture drawn from the other datasets, with probability 0.3.

3.6. Inference and Threshold Selection

Predictions from four rotations are averaged in the original frame to give the test-time-augmented probability map,

$$\hat{P} = \frac{1}{4} \sum_{k=0}^{3} R_{-k}(\sigma(f_\theta(R_k(I)))), \tag{10}$$

where R_k denotes clockwise rotation by $k \times 90^\circ$ and σ is the logistic function.

The binarisation threshold is selected by maximising the Youden J statistic,

$$\tau^* = \arg\max_{\tau} [\mathrm{TPR}(\tau) - \mathrm{FPR}(\tau)], \tag{11}$$

evaluated on the training folds only (the median of the per-image Youden-optimal thresholds of the training-fold images) and applied unchanged to the held-out fold, so that no held-out annotation influences the reported score. Dice, IoU, precision, sensitivity, specificity, clDice and the Hausdorff distance are all computed on the same post-processed binary mask.

After morphological closing with a disk of radius one pixel, connected components smaller than a density-dependent threshold,

$$s_{\min} = \max(16,\ \lfloor 128(1 - 10\rho) \rfloor), \tag{12}$$

are removed, where ρ is the vessel density of the prediction itself, its skeleton length divided by the image area; no annotation of the held-out image is used. The threshold adapts to the image rather than being fixed for the dataset, which prevents the removal of thin capillaries in sparse images while still suppressing small false-positive blobs in dense ones. Fold-wise fitted thresholds ranged from 0.520 to 0.535. Section 4.1 reports the sensitivity of the result to this choice. Selecting the threshold per image from that image's own annotation, a common practice in the segmentation literature, raises Dice by 0.010 and is reported as an optimistic reference rather than as a held-out result.

Before each evaluation, two unlabelled passes over the validation set update the instance-normalisation statistics from the target distribution, following Schneider et al. [68]. This is an unlabelled test-time adaptation step and should be distinguished from a fully frozen held-out evaluation.

3.7. Cross-Validation, Baselines and Ablation

Five-fold stratified cross-validation is used for the VessMAP experiments, with images allocated by contrast quartile so that each fold spans the quality range. The reduced multi-dataset ablation on DRIVE and CHASE_DB1 uses three-fold cross-validation as specified in Section 3.8. Reported metrics are Dice, intersection over union (IoU), precision, clDice [17], sensitivity (recall), specificity and the 95th-percentile Hausdorff distance. clDice and the Hausdorff distance are the topology- and boundary-sensitive measures; IoU and precision are computed per image from the same binary masks. Significance between paired per-image results is assessed by the Wilcoxon signed-rank test with Cohen's d for paired samples (d_z, the mean paired difference divided by its standard deviation).

Two families of comparison are reported. Three unpreprocessed baselines (ResNet-34 U-Net, ResNet-18 U-Net and EfficientNet-B0 U-Net) were trained on raw images, and, therefore, measure the combined contribution of the preprocessing chain and the architectural additions. Three preprocessed architectures (a cascaded two-level minimalist network in the W-Net family [22] with 59,034 parameters, and the proposed pipeline on EfficientNet-B0 and EfficientNet-B4) were all trained with the same reduced chain and, therefore, isolate architecture alone.

Ten preprocessing conditions were evaluated under the full five-fold protocol with a fixed seed. Condition A0 ran the complete seven-stage chain as reference; A1 to A7 removed one stage each; A8 retained intensity normalisation and gamma calibration only; A9 retained intensity normalisation and adaptive CLAHE only. A full factorial design over seven binary stages would require $2^7 = 128$ conditions, or 640 training runs under five-fold cross-validation. The ten conditions instead follow a leave-one-stage-out

design, which measures each stage's marginal contribution in the context of all the others, complemented by two minimal chains that test the opposite extreme of retaining only the always-on normalisation and one further stage. Higher-order interactions are not estimated exhaustively; Section 4.2 checks instead whether the single-stage effects add up to the effect of the reduced chains. Each condition is compared with the complete chain by a paired Wilcoxon test over the 100 per-image results, with Holm correction across the nine comparisons. To keep the ten conditions comparable at modest cost, ablation models are evaluated at a fixed threshold of 0.5 without the test-time normalisation update, so their absolute Dice is lower than that of Section 4.1; only differences between conditions are interpreted.

Inter-annotator agreement was estimated on the 20 dual-annotated images with a bootstrap 95% confidence interval over 10,000 resamples. The representativeness of that subset was checked by comparing its distributions of the four quality measures distributed with VessMAP [41] against the remaining 80 images.

3.8. Multi-Dataset Ablation

To test whether the ablation verdict is a property of the modality or of the particular dataset, a reduced condition set (A0, A1, A4, A6, A8, A9) was repeated on DRIVE and CHASE_DB1, and on VessMAP for comparison, under three-fold cross-validation. Gate thresholds, stratification bins and Frangi ridge polarity were recomputed independently for each dataset, so the self-calibrating mechanism operated separately in each case (dark-ridge Frangi filtering on the two fundus datasets). Training used the VessMAP hyperparameters and the EfficientNet-B4 backbone, with final-epoch weights.

3.9. Cross-Dataset Experimental Protocol

Four LODO experiments were run, each withholding one dataset and training on the other three, with per-dataset thresholds computed independently; each model was trained for 30 epochs with learning-rate reduction on plateau of the training loss and evaluated with its final-epoch weights. LODO and cross-modality models use the U-Net++ architecture of Section 3.5 with an EfficientNet-B2 encoder in place of EfficientNet-B4, to shorten training. The cross-modality direction was additionally run under three resampling regimes, because upsampling acts as a low-pass filter while stage 4 is gated on high-frequency noise: measured high-frequency energy in VessMAP falls by 68% under twofold upsampling. The regimes were upsampling VessMAP to 512×512 while downsampling the fundus sets, resampling all datasets to VessMAP's native 256×256 grid with area averaging, and training on native-resolution crops.

Checkpoints for the cross-modality experiment were selected on a held-out split of the training domains, not on the target set: 10 of the 68 fundus training images were held out as source-domain validation data. Cross-modality models were trained for 30 epochs with Adam at learning rate 3×10^{-4}, batch size 4, seed 42 and learning-rate reduction on plateau of the source-validation Dice. Target-domain Dice was logged every five epochs for analysis only and never used for model selection, because selecting on it would amount to early stopping on the test data.

4. Results

4.1. Cross-Validation Performance

Five-fold cross-validation on VessMAP gave Dice 0.882 ± 0.006 under the training-fold threshold protocol, corresponding to a mean IoU of 0.792 ± 0.009. Table 3 shows that this figure is insensitive to threshold selection. This evaluation uses the reduced chain of condition A9 with test-time normalisation and training-fold thresholds; the stage ablation

of Table 4 uses the simpler fixed-threshold protocol of Section 3.7, under which the same A9 chain scores 0.875. A single threshold fitted to all 100 images pooled gives 0.882 and the conventional 0.5 gives 0.878, while the per-image oracle rule gives 0.892. The training-fold and fixed rules, which use no held-out annotation, agree to within 0.004, and the mean per-image oracle threshold was 0.525, close to the conventional default. The oracle rule raises per-image Dice by 0.010 on average (Wilcoxon $p = 1.2 \times 10^{-4}$, $d_z = 0.44$, $n = 100$).

Table 3. Dice under four rules for choosing the binarisation threshold, applied to identical probability maps from the same trained models. Mean ± standard deviation across folds.

Threshold Rule	Dice	Uses Held-Out Labels
Per-image, fitted on the test image	0.892 ± 0.006	Yes
Fitted on training folds	0.882 ± 0.006	No
Single threshold, all images pooled	0.882 ± 0.005	Yes (pooled)
Fixed at 0.5	0.878 ± 0.006	No

Inter-annotator agreement on the dual-annotated subset is 0.878, with a bootstrap 95% confidence interval of $[0.864, 0.891]$. The cross-validation mean falls inside this interval, so the method matches annotator-level agreement rather than exceeding it (Figure 2a). Sixty-two of 100 images exceed the agreement point estimate and 43 exceed the interval's upper bound. Exceedance rises with image contrast: 12, 16, 16 and 18 of the 25 images in each contrast quartile exceed the point estimate (48% in the lowest against 72% in the highest), and 6, 11, 12 and 14 exceed the upper bound (Figure 2b), so the aggregate count is weighted toward the easier images.

The dual-annotated subset spans 77% to 94% of the full dataset's range on each of the four VessMAP quality measures, with no feature differing detectably between subset and remainder (Mann–Whitney $p \geq 0.08$).

Precision, sensitivity, specificity, IoU, clDice and the 95th-percentile Hausdorff distance for the reduced chain are reported per architecture in Section 4.4.

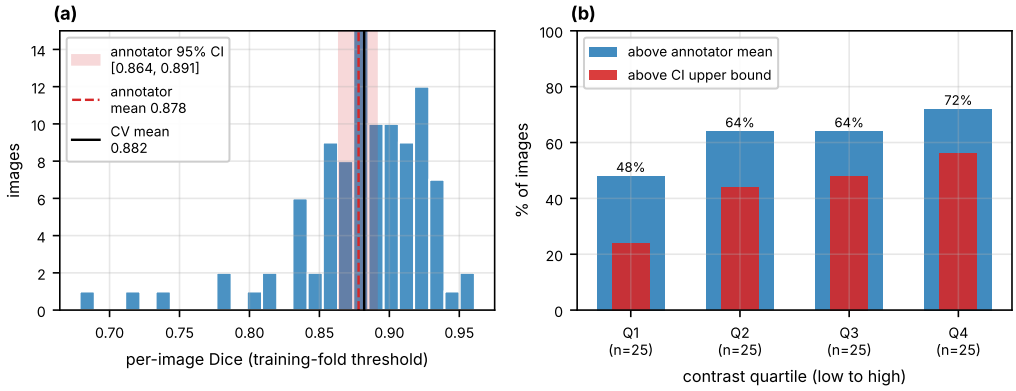


Figure 2. Segmentation performance relative to inter-annotator agreement. (**a**) Distribution of per-image Dice across the 100 VessMAP images under the training-fold threshold rule (the run of Table 3), with the cross-validation mean of 0.882 marked; blue bars give the number of images in each Dice bin. The shaded band is the bootstrap 95% confidence interval of agreement between the two annotators on the dual-annotated subset, $[0.864, 0.891]$; the dashed line is its mean. (**b**) Percentage of images exceeding the annotator mean and the upper bound of its confidence interval, stratified by image contrast quartile ($n = 25$ images per quartile).

4.2. Ablation on Cortical Images

Table 4 reports the ten preprocessing conditions. Removing intensity normalisation costs 0.017 Dice ($p_{\text{Holm}} < 10^{-4}$), the only significant negative delta in the table. It is the

only stage that runs on every image in every condition, and the only one that addresses the absolute pixel-value range the encoder ultimately sees.

Table 4. Per-stage ablation on VessMAP, five-fold cross-validation (mean ± standard deviation across folds; fixed threshold of 0.5, see Section 3.7). ΔDice is condition mean minus the complete-chain mean. A positive delta means the removed stage was harmful. p_{Holm}: paired Wilcoxon test of per-image Dice against the complete chain ($n = 100$), Holm-corrected over nine comparisons.

ID	Condition	Dice	ΔDice	p_{Holm}	IoU	clDice	Sens.
A0	Complete chain	0.850 ± 0.014	—	—	0.744	0.902 ± 0.011	0.933
A1	No intensity normalisation	0.834 ± 0.017	-0.017	$<10^{-4}$	0.720	0.893 ± 0.012	0.935
A2	No homomorphic correction	0.854 ± 0.012	$+0.003$	0.031	0.749	0.905 ± 0.014	0.937
A3	No adaptive CLAHE	0.849 ± 0.012	-0.001	0.31	0.742	0.900 ± 0.007	0.934
A4	No denoising	0.857 ± 0.009	$+0.006$	0.26	0.753	0.913 ± 0.010	0.940
A5	No Frangi blending	0.855 ± 0.013	$+0.004$	$<10^{-4}$	0.750	0.907 ± 0.013	0.935
A6	No LCN	0.870 ± 0.008	$+0.019$	$<10^{-4}$	0.772	0.930 ± 0.011	0.941
A7	No gamma calibration	0.853 ± 0.009	$+0.002$	0.21	0.748	0.901 ± 0.010	0.933
A8	Stages 1 and 7 only	0.877 ± 0.006	$+0.027$	$<10^{-4}$	0.784	0.936 ± 0.008	0.943
A9	Stages 1 and 3 only	0.875 ± 0.009	$+0.025$	$<10^{-4}$	0.781	0.933 ± 0.007	0.944

Figure 3 shows the same conditions graphically. The largest positive single-stage delta is removing LCN, at +0.019 Dice. It is accompanied by a rise in clDice from 0.902 to 0.930, the largest topological gain among single-stage removals. Removing Frangi blending (+0.004) and homomorphic correction (+0.003) also raises Dice significantly, although by much less, while the removal effects of denoising, CLAHE and gamma calibration are not significant after correction. Removing denoising (A4) mainly raises sensitivity (+0.007, $p < 10^{-8}$), consistent with the denoising stage thinning the vessels, whereas removing LCN (A6) mainly raises precision (+0.026, against +0.008 for sensitivity), consistent with the noise amplification discussed in Section 5.

The two reduced chains clear the complete chain by 0.025 to 0.027 Dice ($p_{\text{Holm}} < 10^{-4}$), about twice the inter-fold standard deviation, while the difference between them is 0.002; A8 led A9 by 0.002 to 0.004 in each repeated run, less than the inter-fold standard deviation. The choice between the reduced variants is, therefore, not resolved on VessMAP alone; A9 is used as the reduced chain because it matched or exceeded A8 on the fundus datasets and was the best single chain across the three datasets (Section 4.3).

The single-stage effects do not add up to the reduced-chain effects. Condition A8 removes stages 2 to 6, whose individual removal deltas sum to +0.032 (from unrounded values), yet A8 gains +0.027; condition A9 removes stages 2, 4, 5, 6 and 7, whose deltas sum to +0.035, yet A9 gains +0.025. Removing local contrast normalisation alone (+0.019) accounts for about three quarters of the gain of either reduced chain. The remainder comes from several small effects, of which only Frangi blending and homomorphic correction are individually significant. The complete chain is, therefore, worse than the reduced chains mainly because of one harmful stage, with a smaller contribution from two others.

The Frangi blend weight α, the one parameter not derived from percentile statistics, was swept over $\{0, 0.10, 0.25, 0.50, 0.75, 1.00\}$ within the complete chain, in a separate five-fold run scored with the per-image threshold rule, so absolute values exceed those of Table 4. Dice was flat up to $\alpha = 0.25$ (0.870, 0.871 and 0.868 at $\alpha = 0$, 0.10 and 0.25, within the inter-fold standard deviation of 0.012) and declined at larger weights, to 0.838 at $\alpha = 1.0$; a contrast-calibrated adaptive rule gave 0.868. Blending did not improve segmentation beyond fold noise at any weight.

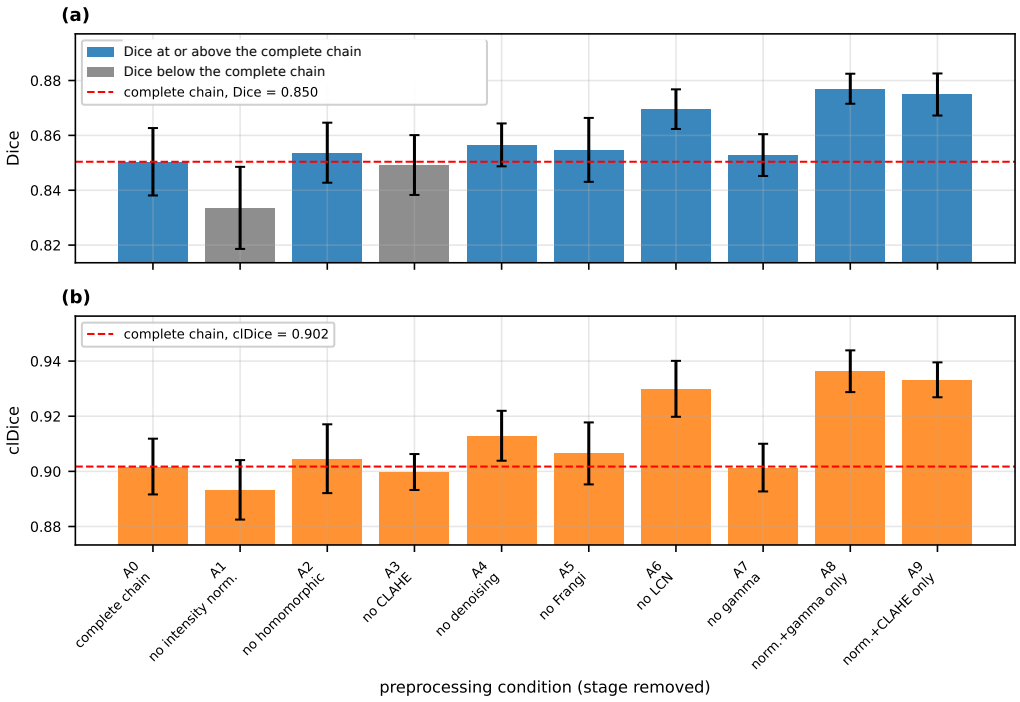


Figure 3. Per-stage ablation of the preprocessing chain on VessMAP, five-fold cross-validation (bars are fold means, error bars the standard deviation across folds). (**a**) Dice and (**b**) centreline Dice for the complete chain and for each condition with one stage removed or reduced to two stages. The dashed line marks the complete chain. Orange bars in panel b show centreline Dice. Bar shading distinguishes conditions whose mean Dice equals or exceeds that of the complete chain from those that fall below it; Table 4 gives the paired tests, under which only the reduction for A1 is significant. In (**b**), removing local contrast normalisation raises centreline Dice as well as Dice, which identifies that stage as breaking vessel connectivity rather than merely shifting boundaries.

4.3. Stage Effects on the Fundus Datasets

Repeating the reduced condition set on DRIVE and CHASE_DB1, with VessMAP repeated under the same three-fold protocol for comparison, gives Table 5. The stages whose removal helps on VessMAP also help when removed from the fundus chains. Removing local contrast normalisation raises Dice by 0.012, 0.007 and 0.017 on VessMAP, DRIVE and CHASE_DB1, removing denoising by 0.007, 0.038 and 0.012, and both two-stage chains exceed the complete chain on every dataset, by 0.023 to 0.056. Intensity normalisation is the only stage whose effect differs in sign: its removal costs 0.015 on VessMAP and 0.035 on DRIVE but changes CHASE_DB1 by +0.003, inside that dataset's fold variability.

Table 5. Effect of removing each stage, and of reducing the chain to two stages, expressed as ΔDice against the complete chain on one cortical confocal dataset and two retinal fundus datasets (three-fold cross-validation, EfficientNet-B4, VessMAP hyperparameters). A positive value means the change improved segmentation. Values in parentheses are the fold-level standard deviation of the condition.

Change	VessMAP (Confocal)	DRIVE (Fundus)	CHASE_DB1 (Fundus)
No intensity normalisation (A1)	−0.015 (0.011)	−0.035 (0.026)	+0.003 (0.005)
No denoising (A4)	+0.007 (0.007)	+0.038 (0.027)	+0.012 (0.014)
No local contrast normalisation (A6)	+0.012 (0.002)	+0.007 (0.001)	+0.017 (0.013)
Stages 1 and 7 only (A8)	+0.026 (0.003)	+0.055 (0.021)	+0.039 (0.017)
Stages 1 and 3 only (A9)	+0.023 (0.004)	+0.055 (0.004)	+0.056 (0.017)
Complete-chain Dice	0.853 ± 0.009	0.667 ± 0.008	0.680 ± 0.003

With three folds on 20 and 28 images, the fundus deltas are imprecise, and the DRIVE and CHASE_DB1 effects of single-stage removal lie within one or two fold standard deviations. None of them, however, run opposite to the VessMAP result for local contrast normalisation, denoising or the reduced chains, and the reduced-chain gains on the fundus datasets exceed their fold variability.

Per-dataset chain selection was evaluated as a candidate automatic configuration mechanism. It gave a mean Dice of 0.779 across the three datasets, against 0.778 for the best single fixed chain (A9) applied everywhere and 0.733 for the complete chain, and importing another dataset's selected chain cost 0.004 Dice on average. A single reduced chain is, therefore, an adequate recommendation across these datasets.

Absolute Dice on DRIVE and CHASE_DB1 (0.667 and 0.680) is below published fundus-domain results, because these models were trained under hyperparameters selected for VessMAP with three folds on 20 and 28 images respectively. The per-stage deltas, not the absolute values, are the quantity of interest here.

4.4. Architecture

Table 6 and Figure 4 compare architectures. Among models sharing the same reduced preprocessing chain, a minimalist network with 59,034 parameters reaches a Dice of 0.878, within 0.012 of a pipeline with approximately 354 times more parameters ($\Delta = 0.012$, $p = 3.7 \times 10^{-12}$, $d_z = 0.80$, $n = 100$). The difference is statistically detectable but small relative to the inter-annotator interval of Section 4.1. Precision is about 0.10 below sensitivity for every preprocessed model, the expected consequence of a Tversky weighting that penalises missed vessels more than false detections, and the W-Net trails both EfficientNet models under IoU and precision as under Dice. On clDice the EfficientNet-B4 model exceeds B0 by 0.002 ($p = 0.007$, $d_z = 0.28$), while B0 exceeds the W-Net by 0.004 ($p = 0.061$).

Table 6. Architecture comparison on VessMAP, five-fold cross-validation. The upper group was trained on raw images and, therefore, measures the combined contribution of preprocessing and architecture; the lower group shares the same reduced preprocessing chain and, therefore, isolates architecture. Dice in this table uses the per-image oracle threshold, and these models were trained in runs separate from those of Table 3; see Table 3, which shows this rule to exceed the training-fold rule by about 0.010 for the selected model. Dice is the mean ± standard deviation across folds; the other columns are means across folds, with fold-level standard deviations of at most 0.010 for IoU, precision, sensitivity, specificity and clDice. IoU and precision are computed per image from the same binary masks as Dice; HD95 is the 95th-percentile Hausdorff distance in pixels.

Model	Params	Dice	IoU	Prec.	Sens.	Spec.	clDice	HD95
Inter-annotator agreement [1]	—	0.878	—	—	—	—	—	—
Trained on raw images								
ResNet-18 U-Net	11 M	0.865 ± 0.008	—	—	—	—	—	—
ResNet-34 U-Net	21 M	0.868 ± 0.009	—	—	—	—	—	—
EfficientNet-B0 U-Net	6.6 M	0.868 ± 0.010	—	—	—	—	—	—
Trained with the reduced preprocessing chain								
W-Net (minimalist)	59 K	0.878 ± 0.006	0.784	0.830	0.933	0.933	0.937	2.24
Proposed, EfficientNet-B0	6.6 M	0.889 ± 0.005	0.802	0.840	0.946	0.938	0.941	2.71
Proposed, EfficientNet-B4	20.9 M	0.889 ± 0.006	0.802	0.840	0.947	0.938	0.943	2.73

[1] Estimated on the 20-image dual-annotated subset; bootstrap 95% confidence interval $[0.864, 0.891]$. Per-image metrics other than Dice were not retained for the raw-image baselines.

Holding every other component fixed and varying only the encoder, EfficientNet-B0 (6.6 M parameters) and EfficientNet-B4 (20.9 M) differ by 0.0004 Dice ($p = 0.33$, $d_z = 0.07$, $n = 100$). A three-fold increase in encoder capacity, therefore, yields no detectable difference

in Dice, and the generalisation gap, measured as the mean difference between training and validation Dice over the final five epochs, was −0.052 for EfficientNet-B0 and −0.049 for EfficientNet-B4; the gap is negative because training Dice is computed on augmented batches with the network in training mode. Encoder capacity is not the limiting factor on this dataset, and EfficientNet-B0 is recommended for further use on efficiency grounds.

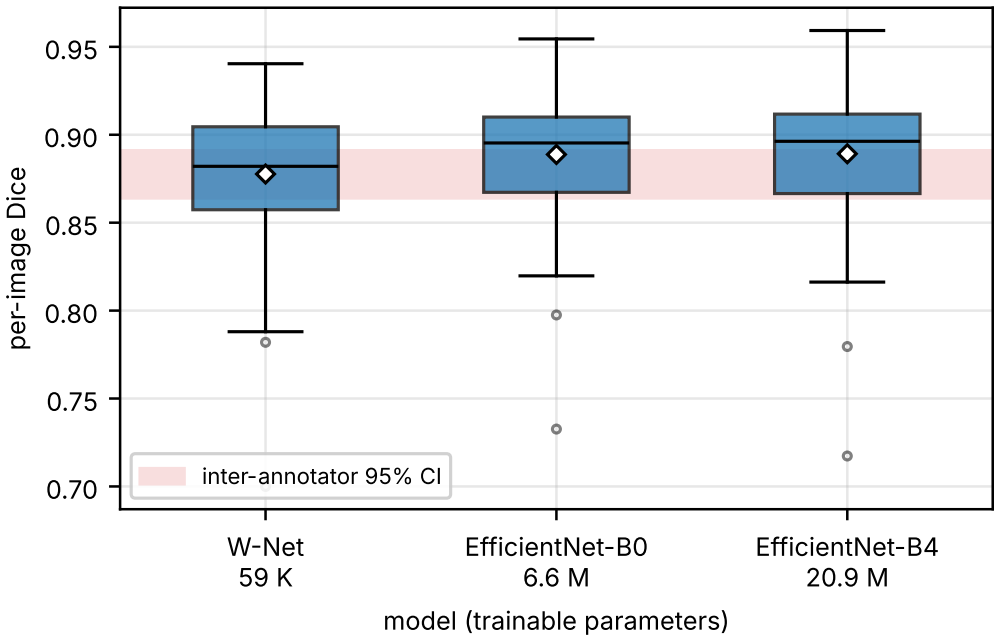


Figure 4. Per-image Dice on VessMAP for segmentation models spanning approximately 2.55 orders of magnitude in parameter count, all trained with the same preprocessing chain, loss, augmentation, schedule and five-fold split (the runs of Table 6, per-image oracle threshold). Boxes show the interquartile range, the central line the median, and whiskers 1.5 times the interquartile range; diamonds mark the means, and small circles mark images outside the whiskers. The shaded band is the 95% confidence interval of inter-annotator agreement.

4.5. Per-Image Failure Analysis

Errors were characterised from the held-out per-image metrics of the A9 chain, the chain used in the main run, in the ablation of Table 4 (fixed threshold of 0.5). Across the 100 images, Dice tracked precision (Spearman $\rho = 0.82$, $p < 10^{-25}$) but not sensitivity ($\rho = -0.11$, $p = 0.26$). In the ten lowest-Dice images (mean Dice 0.768), sensitivity stayed high (0.948) while precision fell to 0.658, and precision was below sensitivity in nine of the ten; in the ten highest-Dice images, precision was 0.916 and sensitivity 0.959. The dominant failure mode is, therefore, over-segmentation, that is, false-positive response in regions the annotator marked as background, rather than missed vessels. This matches the gap between mean sensitivity (0.944) and mean precision (0.823) over all images, and the precision–sensitivity imbalance of all three architectures in Table 6.

4.6. Cross-Dataset Generalisation

The three fundus LODO experiments gave Dice of 0.629 (DRIVE), 0.642 (STARE) and 0.630 (CHASE_DB1), a mean of 0.634 ± 0.008. Variance across datasets that differ in scanner, field of view and patient population is low. Training on the three fundus datasets and testing on VessMAP gave 0.099, a gap of 0.535 below the fundus mean and a qualitatively different regime.

Because the naive protocol resamples VessMAP upward while resampling the fundus sets downward, the cross-modality experiment was repeated under three regimes

(Figure 5a). In this separate experiment, at a fixed threshold of 0.5, Dice was 0.086 with VessMAP upsampled, 0.096 on the common 256 × 256 grid, and 0.039 on native-resolution crops. The difference between the first two is small and did not hold a consistent sign across repeated runs; the measured transfer is, therefore, sensitive to both the resampling regime and to training noise. The native-crop regime performs far worse, consistent with a vessel-calibre mismatch, since fundus vessels occupy roughly 10 to 15 pixels at native resolution against 2 to 5 in VessMAP.

Mean per-image area under the receiver operating characteristic curve was 0.473 (Figure 5b), with 79% of test images below 0.5 and mean predicted probability lower inside annotated vessels (0.171) than outside them (0.211). Fitting a threshold to each test image's own annotation raises Dice from 0.096 to 0.337 without any change in ranking. Neither vessel density nor contrast was associated with the per-image AUC (Spearman $\rho = 0.18$, $p = 0.07$, and $\rho = -0.01$, $p = 0.92$).

Figure 5c shows the training curves of the same run in the common-grid regime. Source-domain validation Dice rose from 0.26 after the first epoch (epoch 0) to 0.54 at the last (epoch 29), whereas target-domain Dice was highest after the first epoch (0.23), fell to 0.12 by epoch 5 and stayed between 0.11 and 0.14 thereafter. The decline is not monotonic, but target-domain Dice was likewise highest at epoch 0 in the other two resampling regimes of this run and in all three regimes of three earlier runs, while source-domain validation Dice rose. Improvement on the fundus domain did not carry over to the cortical domain.

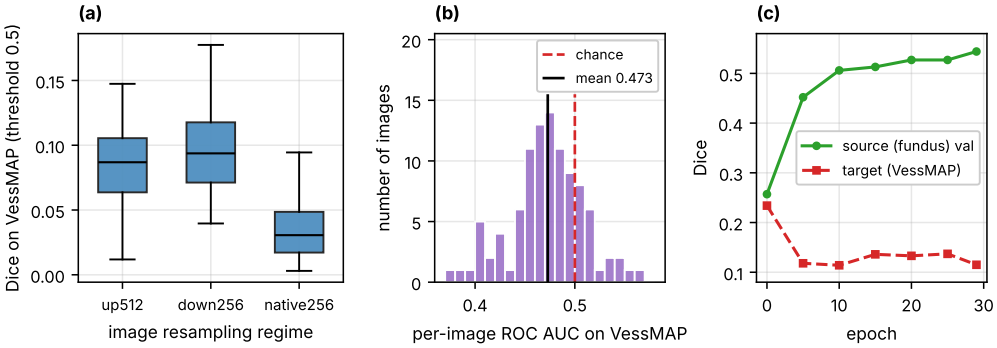


Figure 5. Cross-modality transfer from retinal fundus photography to cortical confocal fluorescence microscopy. Models were trained on DRIVE, CHASE_DB1 and STARE and tested on VessMAP. (**a**) Dice at a fixed threshold of 0.5 under three image-resampling regimes; blue boxes show the interquartile range and median, and whiskers extend to 1.5 times the interquartile range. (**b**) Distribution of per-image area under the receiver operating characteristic curve in the common-grid regime (purple bars), with chance level (dashed red line) and the mean (solid black line); values near 0.5 indicate that the ranking of vessel against background pixels is not preserved across modalities. (**c**) Dice on held-out source-domain images and on the target domain during training (solid green and dashed red lines), in the common-grid regime of the same run, logged every five epochs over the 30-epoch schedule of Section 3.9.

5. Discussion

Several of the imported stages perform poorly on VessMAP, plausibly because their assumptions were developed around the image-quality characteristics of fundus photography, whereas VessMAP differs substantially in contrast, density and noise. Local contrast normalisation is the clearest case. It divides by a local standard deviation that collapses toward zero across the dark regions that dominate a confocal field, so the normalisation amplifies background noise to the order of the vessel signal; removing it recovered 0.019 Dice and 0.028 clDice, the largest topological gain among single-stage removals. Bilateral and wavelet denoising point the same way, though their effect was not significant, and for the

same reason: vessel calibre in VessMAP is two to five pixels and the intensity transition at the wall spans one to three, placing the boundary inside the band that soft-thresholding targets, so the predictions thinned rather than cleaned. Frangi blending slightly reduced accuracy (removal +0.004 Dice), consistent with a vesselness prior being redundant for a pretrained encoder, and gamma calibration produced little additional change, potentially because the first convolution applies instance normalisation. The multi-dataset ablation shows, however, that the harm is not confined to the modality change. Removing local contrast normalisation or denoising, and reducing the chain to two stages, also improved Dice on DRIVE and CHASE_DB1, the fundus datasets for which these operations were developed. Stacking illumination, contrast and noise corrections in front of a pretrained encoder appears to distort the input more than any single correction helps, at least under the shared training protocol used here, which was tuned on VessMAP and may itself favour shorter chains. Intensity normalisation is the only stage whose usefulness varied: it was required on VessMAP and DRIVE but not on CHASE_DB1.

The additivity analysis of Section 4.2 sharpens the answer to why the complete chain underperforms. About three quarters of its deficit comes from one stage, local contrast normalisation, with small but significant contributions from Frangi blending and homomorphic correction; the remaining imported stages are individually neutral on this dataset. This argues for using such a chain as a set of candidate operations to be pruned by ablation on the target data, rather than as a fixed multi-stage pipeline: in this framing the useful product of the method is the removal of operations that harm the target data, and the percentile gates serve only to apply the retained operations where the image warrants them.

Two consequences follow. First, architecture is not the constraint on this dataset. A 59,000-parameter network came within 0.012 Dice of a pipeline 354 times larger, tripling encoder capacity produced no detectable change in Dice, and the generalisation gap was similar for both encoders; preprocessing and inference protocol account for more of the achievable accuracy than depth does, and the smaller encoder should be preferred. Second, transfer across this modality boundary fails at the level of representation rather than calibration. Mean per-image ROC AUC was 0.473, mean predicted probability was lower inside annotated vessels than outside them, and 79% of test images had a per-image AUC below 0.5, so thresholding cannot separate vessels from background on those images. Vessel polarity is the most likely cause, since retinal vessels are dark in the green channel used for training while cortical vessels are bright, and shape-prior pre-training transfers to VessMAP without target-domain data [69], which is consistent with texture-derived features being what fails to cross. Neither vessel density nor contrast predicted the per-image ranking failure. The mechanism proposed elsewhere for such collapses, confident over-prediction under a higher-density prior, is not supported: no test image met the background false-positive criterion, and the observed failure is under-confidence. A direct numerical comparison against the retinal literature would in any case be misleading, since a Dice obtained on DRIVE and one obtained on VessMAP are not measurements of the same quantity [70,71]; comparable two-dimensional results on VessMAP itself have only recently begun to appear [69]. The practical implication is deliberately narrow. A preprocessing chain should not be assumed to be portable across vascular imaging domains because its stages encode assumptions about contrast, density and polarity. Stage-wise evaluation on target data provides a direct way to identify which operations remain useful after transfer. These are observations on one cortical confocal dataset and two fundus datasets; they show that stages imported from fundus pipelines can reduce accuracy on cortical and fundus data alike, so that their usefulness has to be established on the target dataset, not that any particular stage is beneficial or harmful for vascular imaging in general.

Limitations

VessMAP comprises 100 images from a single laboratory and a single acquisition protocol. The reduced chain was identified by ablation on those images, and a laboratory using a different confocal microscope, fluorophore or illumination wavelength might find that different stages are useful.

The ablation was conducted at a fixed random seed. The four effects that survive Holm correction (intensity normalisation, local contrast normalisation, Frangi blending and homomorphic correction) are supported by paired per-image tests. A repeat run of the ablation reproduced the first three (-0.032, $+0.015$ and $+0.003$) and the gains of the reduced chains ($+0.019$ to $+0.023$), but not the small homomorphic effect (-0.007), which should, therefore, not be relied on; the two reduced chains differ by less than the inter-fold standard deviation.

The multi-dataset comparison used three folds and hyperparameters selected for VessMAP, so the absolute fundus Dice values are not competitive with the fundus-domain literature. Only the per-stage deltas are interpreted, and the shared hyperparameters may themselves favour shorter chains on the fundus data.

The cross-modality experiment was run at a single seed, and repeated runs showed that the sign of the difference between resampling regimes was not stable. The reported test-time protocol also updates instance-normalisation statistics from unlabelled validation images; because this uses target-distribution information, its relationship to a strictly held-out evaluation should be stated explicitly and should not be presented as equivalent to a fully frozen test protocol.

The headline results use thresholds fitted on training folds; the stage ablation (Table 4) uses a fixed threshold of 0.5, and the architecture comparison (Table 6) and the Frangi sweep use the per-image rule. Absolute Dice under a per-image rule fitted to test annotations is approximately 0.010 higher; within each table the same rule applies to every condition, so the paired differences are unaffected. Three runs of the main cross-validation under the final protocol gave training-fold-threshold Dice of 0.882, 0.877 and 0.878; the first, whose per-image outputs accompany the ablation, is reported. Four earlier runs under a protocol that selected checkpoints on held-out folds and gated stages on annotation-derived quality measures gave 0.874 to 0.881.

The Tversky weights were set from a dataset-level density averaged over all 100 VessMAP images, including those later held out in each fold. The resulting scalar affects only the loss weighting and differs from the low-density default ($\alpha = 0.15$) by 0.005, but it is a dataset constant rather than a per-fold training statistic.

The inter-annotator agreement estimate rests on 20 of 100 images. Although that subset spans most of the dataset's quality range, an interval of this width cannot distinguish a method matching annotator agreement from one modestly above or below it.

An explainability analysis relating per-image performance to image quality features was carried out but is not reported as a finding, because the feature ranking did not replicate across repeated runs of the full pipeline.

6. Conclusions

The experiments evaluated whether a preprocessing chain assembled from the retinal fundus literature remains appropriate for confocal fluorescence microscopy of the mouse cortex. A five-fold ablation of its seven stages showed that several imported stages do not transfer unchanged: only intensity normalisation proved indispensable, its removal costing 0.017 Dice, while reduced two-stage chains outperformed the complete chain by 0.025 to 0.027 Dice. In a matched three-fold ablation, removing local contrast normalisation or denoising, and reducing the chain to two stages, improved Dice on DRIVE and CHASE_DB1

as well, so the imported chain was not the best configuration even for the fundus images it was drawn from. The resulting pipeline reached a cross-validation Dice of 0.882 ± 0.006 under a threshold protocol whose fitting uses no held-out annotations, within the inter-annotator agreement interval of $[0.864, 0.891]$. A 59,034-parameter network came within 0.012 Dice of the full pipeline and tripling encoder capacity produced no detectable change, while transfer from fundus training gave a mean per-image ROC AUC of 0.473, indicating that vessel-versus-background ranking is not preserved across the modality boundary.

On the datasets examined, preprocessing operations established on colour fundus photography did not transfer unchanged to cortical fluorescence microscopy, and the complete chain underperformed reduced chains on the fundus datasets too. The results support stage-wise validation on the target modality and dataset before adopting such a chain. Two directions follow. Extending the protocol to further modalities would establish whether image properties such as polarity and vessel density suffice to predict a stage's contribution in advance. It also remains open whether ablation-selected chains are complementary to shape-prior pre-training [69] and to label-free domain adaptation.

Author Contributions: Conceptualization, S.A.; methodology, S.A.; software, S.A.; validation, S.A.; formal analysis, S.A.; investigation, H.T.S.; writing—original draft preparation, S.A.; writing—review and editing, H.T.S.; visualization, H.T.S.; supervision, H.T.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable. This study is a secondary computational analysis of publicly available datasets; no experiments on humans or animals were conducted and no new human or animal data were collected.

Informed Consent Statement: Not applicable.

Data Availability Statement: The VessMAP dataset analysed in this study is publicly available at https://zenodo.org/records/10045265 (accessed on 6 September 2026) under a Creative Commons Attribution licence, as reported in its source publication [41]. DRIVE, STARE and CHASE_DB1 are available from their respective sources. No new data were generated in this study. The analysis notebook, including switches that reproduce earlier protocol variants, is available from the corresponding author on reasonable request.

Acknowledgments: The authors thank the creators of VessMAP for releasing their dataset publicly with dual annotations, which made the inter-annotator agreement analysis possible. Experiments were carried out using GPU compute resources provided by the Kaggle platform. The authors used Claude (Anthropic) to assist with debugging Python 3.12 runtime errors during code development, with language editing, and with producing the block diagram in Figure 1. Experimental design, hypothesis formulation, data analysis, interpretation of results and the writing of this manuscript are the authors' own work.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Blinder, P.; Tsai, P.; Kaufhold, J.; Knutsen, P.; Suhl, H.; Kleinfeld, D. The cortical angiome: An interconnected vascular network with noncolumnar patterns of blood flow. *Nat. Neurosci.* **2013**, *16*, 889–897. [CrossRef] [PubMed]
2. Pian, Q.; Alfadhel, M.; Tang, J.; Lee, G.; Li, B.; Fu, B.; Ayata, Y.; Yaseen, M.; Boas, D.; Secomb, T.; et al. Cortical microvascular blood flow velocity mapping by combining dynamic light scattering optical coherence tomography and two-photon microscopy. *J. Biomed. Opt.* **2023**, *28*, 076003. [CrossRef] [PubMed]
3. Soares, J.; Leandro, J.; Cesar, R.; Jelinek, H.; Cree, M. Retinal vessel segmentation using the 2-D Gabor wavelet and supervised classification. *IEEE Trans. Med. Imaging* **2006**, *25*, 1214–1222. [CrossRef] [PubMed]
4. Fraz, M.; Remagnino, P.; Hoppe, A.; Uyyanonvara, B.; Rudnicka, A.; Owen, C.; Barman, S. An ensemble classification-based approach applied to retinal blood vessel segmentation. *IEEE Trans. Biomed. Eng.* **2012**, *59*, 2538–2548. [CrossRef] [PubMed]

5. Badawi, S.; Fraz, M. Optimizing the trainable B-COSFIRE filter for retinal blood vessel segmentation. *PeerJ* **2018**, *6*, e5855. [CrossRef] [PubMed]
6. Iqbal, M.; Riaz, M.; Ghafoor, A.; Ahmad, A. A vessel segmentation technique for retinal images. *Int. J. Imaging Syst. Technol.* **2021**, *31*, 160–167. [CrossRef]
7. Zhang, B.; Zhang, L.; Zhang, L.; Karray, F. Retinal vessel extraction by matched filter with first-order derivative of Gaussian. *Comput. Biol. Med.* **2010**, *40*, 438–445. [CrossRef] [PubMed]
8. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of MICCAI*; Springer: Cham, Switzerland, 2015; pp. 234–241.
9. Zhou, Z.; Rahman Siddiquee, M.; Tajbakhsh, N.; Liang, J. UNet++: A nested U-Net architecture for medical image segmentation. In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*; Springer: Cham, Switzerland, 2018; pp. 3–11.
10. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In *Proceedings of CVPR*; IEEE: Piscataway, NJ, USA, 2018; pp. 7132–7141.
11. Roy, A.; Navab, N.; Wachinger, C. Recalibrating fully convolutional networks with spatial and channel squeeze and excitation blocks. *IEEE Trans. Med. Imaging* **2019**, *38*, 540–549. [CrossRef] [PubMed]
12. Wang, D.; Haytham, A.; Pottenburgh, J.; Saeedi, O.; Tao, Y. Hard attention net for automatic retinal vessel segmentation. *IEEE J. Biomed. Health Inform.* **2020**, *24*, 3384–3396. [CrossRef] [PubMed]
13. Jiang, M.; Zhu, Y.; Zhang, X. CoVi-Net: A hybrid convolutional and vision transformer neural network for retinal vessel segmentation. *Comput. Biol. Med.* **2024**, *170*, 108047. [CrossRef] [PubMed]
14. Yang, R.; Zhang, S. Enhancing retinal vascular structure segmentation in images with a novel design two-path interactive fusion module model. *arXiv* **2024**, arXiv:2403.01362.
15. Shao, D.; Xu, R.; Ma, L.; Yi, S. Tubular-aware Mamba for accurate retinal vessel segmentation: Preserving fine details and topological connectivity. *Multimed. Syst.* **2025**, *31*, 182. [CrossRef]
16. Perera, S.; Erzurumlu, Y.; Gulati, D.; Yilmaz, A. MobileUNETR: A lightweight end-to-end hybrid vision transformer for efficient medical image segmentation. In *Computer Vision—ECCV 2024 Workshops*; Springer: Cham, Switzerland, 2025; pp. 281–299.
17. Shit, S.; Paetzold, J.; Sekuboyina, A.; Ezhov, I.; Unger, A.; Zhylka, A.; Pluim, J.; Bauer, U.; Menze, B. clDice—A novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of CVPR*; IEEE: Piscataway, NJ, USA, 2021; pp. 16560–16569.
18. He, Z.; Chen, F.; Jiang, F.; Liu, W.; Ye, Z. A dual-branch network based on connectivity mask for retinal vessel segmentation. In *Proceedings of ICME*; IEEE: Piscataway, NJ, USA, 2024; pp. 1–6.
19. Xu, P.; Zhao, G.; Liu, J.; Jahanshahi, H.; Tang, Z.; Gong, S. MCG & BA-Net: Retinal vessel segmentation using multiscale context gating and breakpoint attention. *IET Image Process.* **2022**, *16*, 3039–3056. [CrossRef]
20. Zhou, W.; Wang, X.; Yang, X.; Hu, Y.; Yi, Y. Skeleton-guided multi-scale dual-coordinate attention aggregation network for retinal blood vessel segmentation. *Comput. Biol. Med.* **2024**, *181*, 109027. [CrossRef] [PubMed]
21. Huang, S.; Li, J.; Shen, N.; Xu, T. Meta-tubular-net: A robust topology-aware re-weighting network for retinal vessel segmentation. *Biomed. Signal Process. Control* **2024**, *91*, 106060. [CrossRef]
22. Galdran, A.; Anjos, A.; Dolz, J.; Chakor, H.; Lombaert, H.; Ben Ayed, I. State-of-the-art retinal vessel segmentation with minimalistic models. *Sci. Rep.* **2022**, *12*, 6174. Correction in *Sci. Rep.* **2023**, *13*, 12812. https://doi.org/10.1038/s41598-023-40057-0. [CrossRef] [PubMed]
23. Fu, W.; Breininger, K.; Schaffert, R.; Pan, Z.; Maier, A. Keep it simple, scholar: An experimental analysis of few-parameter segmentation networks for retinal vessels in fundus imaging. *Int. J. Comput. Assist. Radiol. Surg.* **2021**, *16*, 967–978. [CrossRef] [PubMed]
24. Zeng, Z.; Liu, J.; Huang, X.; Luo, K.; Yuan, X.; Zhu, Y. Efficient retinal vessel segmentation with 78K parameters. *J. Imaging* **2025**, *11*, 306. [CrossRef] [PubMed]
25. Pizer, S.; Amburn, E.; Austin, J.; Cromartie, R.; Geselowitz, A.; Greer, T.; ter Haar Romeny, B.; Zimmerman, J.; Zuiderveld, K. Adaptive histogram equalization and its variations. *Comput. Vis. Graph. Image Process.* **1987**, *39*, 355–368. [CrossRef]
26. Frangi, A.; Niessen, W.; Vincken, K.; Viergever, M. Multiscale vessel enhancement filtering. In *Proceedings of MICCAI*; Springer: Berlin, Germany, 1998; pp. 130–137.
27. Mallat, S. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Trans. Pattern Anal. Mach. Intell.* **1989**, *11*, 674–693. [CrossRef]
28. Tomasi, C.; Manduchi, R. Bilateral filtering for gray and color images. In *Proceedings of ICCV*; IEEE: Piscataway, NJ, USA, 1998; pp. 839–846.
29. Oppenheim, A.; Schafer, R.; Stockham, T. Nonlinear filtering of multiplied and convolved signals. *IEEE Trans. Audio Electroacoust.* **1968**, *16*, 437–466. [CrossRef]
30. Mathews, M.; Anzar, S.; Krishnan, R.; Panthakkan, A. EfficientNet for retinal blood vessel segmentation. In *Proceedings of ICSPIS*; IEEE: Piscataway, NJ, USA, 2020; pp. 1–4.

31. Cao, J.; Chen, J.; Gu, Y.; Liu, J. MFA-UNet: A vessel segmentation method based on multi-scale feature fusion and attention module. *Front. Neurosci.* **2023**, *17*, 1249331. [CrossRef] [PubMed]
32. Aurangzeb, K.; Aslam, S.; Alhussein, M.; Naqvi, R.A.; Arsalan, M.; Haider, S.I. Contrast enhancement of fundus images by employing modified PSO for improving the performance of deep learning models. *IEEE Access* **2021**, *9*, 47930–47945. [CrossRef]
33. Peng, L.; Lin, L.; Cheng, P.; He, H.; Tang, X. Student becomes decathlon master in retinal vessel segmentation via dual-teacher multi-target domain adaptation. In *Proceedings of MLMI*; Springer: Cham, Switzerland, 2022; pp. 32–42.
34. Hu, D.; Li, H.; Liu, H.; Wang, J.; Yao, X.; Lu, D.; Oguz, I. AdaptDiff: Cross-modality domain adaptation via weak conditional semantic diffusion for retinal vessel segmentation. In *Proceedings of SASHIMI*; Springer: Cham, Switzerland, 2024; pp. 13–23.
35. Hu, D.; Li, H.; Liu, H.; Yao, X.; Wang, J.; Oguz, I. MAP: Domain generalization via meta-learning on anatomy-consistent pseudo-modalities. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2023 Workshops*; Springer: Cham, Switzerland, 2023; pp. 182–192.
36. Hu, D.; Li, H.; Liu, H.; Yao, X.; Wang, J.; Oguz, I. VesselMorph: Domain-generalized retinal vessel segmentation via shape-aware representation. *arXiv* **2023**, arXiv:2307.00240.
37. Wang, S.; Yu, L.; Li, K.; Yang, X.; Fu, C.; Heng, P. DoFE: Domain-oriented feature embedding for generalizable fundus image segmentation on unseen datasets. *IEEE Trans. Med. Imaging* **2020**, *39*, 4237–4248. [CrossRef] [PubMed]
38. Liu, B.; Zhang, Y.; Wang, S.; Li, S.; Hong, J. DGSSA: Domain generalization with structural and stylistic augmentation for retinal vessel segmentation. *Neural Netw.* **2026**, *194*, 108118. [CrossRef] [PubMed]
39. Todorov, M.; Paetzold, J.; Schoppe, O.; Tetteh, G.; Shit, S.; Efremov, V.; Todorov-Völgyi, K.; Düring, M.; Dichgans, M.; Piraud, M.; et al. Machine learning analysis of whole mouse brain vasculature. *Nat. Methods* **2020**, *17*, 442–449. [CrossRef] [PubMed]
40. Tetteh, G.; Efremov, V.; Forkert, N.; Schneider, M.; Kirschke, J.; Weber, B.; Zimmer, C.; Piraud, M.; Menze, B. DeepVesselNet: Vessel segmentation, centerline prediction, and bifurcation detection in 3-D angiographic volumes. *Front. Neurosci.* **2020**, *14*, 592352. [CrossRef] [PubMed]
41. da Silva, M.; de Carvalho Santos, N.; Ouellette, J.; Lacoste, B.; Comin, C. A new dataset for measuring the performance of blood vessel segmentation methods under distribution shifts. *PLoS ONE* **2025**, *20*, e0322048. [CrossRef]
42. Staal, J.; Abràmoff, M.; Niemeijer, M.; Viergever, M.; van Ginneken, B. Ridge-based vessel segmentation in color images of the retina. *IEEE Trans. Med. Imaging* **2004**, *23*, 501–509. [CrossRef] [PubMed]
43. Hoover, A.; Kouznetsova, V.; Goldbaum, M. Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Trans. Med. Imaging* **2000**, *19*, 203–210. [CrossRef] [PubMed]
44. Budai, A.; Bock, R.; Maier, A.; Hornegger, J.; Michelson, G. Robust vessel segmentation in fundus images. *Int. J. Biomed. Imaging* **2013**, *2013*, 154860. [CrossRef] [PubMed]
45. Jin, K.; Huang, X.; Zhou, J.; Li, Y.; Yan, Y.; Sun, Y.; Zhang, Q.; Wang, Y.; Ye, J. FIVES: A fundus image dataset for artificial intelligence based vessel segmentation. *Sci. Data* **2022**, *9*, 475. [CrossRef] [PubMed]
46. Fu, H.; Xu, Y.; Lin, S.; Wong, D.; Liu, J. DeepVessel: Retinal vessel segmentation via deep learning and conditional random field. In *Proceedings of MICCAI*; Springer: Cham, Switzerland, 2016; pp. 132–139.
47. Araújo, R.; Cardoso, J.; Oliveira, H. A single-resolution fully convolutional network for retinal vessel segmentation in raw fundus images. In *Proceedings of ICIAP*; Springer: Cham, Switzerland, 2019; pp. 59–69.
48. Islam, M.T.; Wu, D.-W.; Tang, Q.-Q.; Zhao, K.-Y.; Yin, T.; Li, Y.-F.; Shang, W.-Y.; Liu, J.-Y.; Zhang, H.-X. Dual encoding feature filtering generalized attention UNet for retinal vessel segmentation. *arXiv* **2025**, arXiv:2506.02312.
49. Xin, Q.; Han, L.; Zou, X. MDF-Net: Multiscale deep fusion network and dual threshold decision for retinal vessel segmentation. In *Proceedings of ISPP*; SPIE: Bellingham, WA, USA, 2024; Volume 13180, p. 131806O.
50. Tian, F.; Li, Y.; Wang, J.; Chen, W. Blood vessel segmentation of fundus retinal images based on improved Frangi and mathematical morphology. *Comput. Math. Methods Med.* **2021**, *2021*, 4761517. [CrossRef] [PubMed]
51. Tavakoli, M.; Kalantari, F.; Golestaneh, A. Comparing different preprocessing methods in automated segmentation of retinal vasculature. In *Proceedings of NSS/MIC*; IEEE: Piscataway, NJ, USA, 2017. [CrossRef]
52. Isensee, F.; Jaeger, P.F.; Kohl, S.A.A.; Petersen, J.; Maier-Hein, K.H. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nat. Methods* **2021**, *18*, 203–211. [CrossRef] [PubMed]
53. Fadugba, J.; Köhler, P.; Koch, L.; Manescu, P.; Berens, P. Benchmarking retinal blood vessel segmentation models for cross-dataset and cross-disease generalization. *arXiv* **2024**, arXiv:2406.14994.
54. Hu, D.; Li, H.; Liu, H.; Oguz, I. Domain generalization for retinal vessel segmentation via Hessian-based vector field. *Med. Image Anal.* **2024**, *95*, 103164. [CrossRef] [PubMed]
55. Kuş, Z. UKnow-Net: Knowledge-enhanced U-Net for improved retinal vessel segmentation. *Gazi Univ. J. Sci. Part A* **2024**, *11*, 742–758. [CrossRef]
56. Ben-David, S.; Blitzer, J.; Crammer, K.; Kulesza, A.; Pereira, F.; Vaughan, J.W. A theory of learning from different domains. *Mach. Learn.* **2010**, *79*, 151–175. [CrossRef]

57. Guan, H.; Liu, M. Domain adaptation for medical image analysis: A survey. *IEEE Trans. Biomed. Eng.* **2022**, *69*, 1173–1185. [CrossRef] [PubMed]
58. Bandara, A.M.R.R.; Giragama, P.W.G.R. A retinal image enhancement technique for blood vessel segmentation algorithm. In *Proceedings of ICIIS*; IEEE: Piscataway, NJ, USA, 2017. [CrossRef]
59. Kumar, K.; Samal, D.; Suraj. Automated retinal vessel segmentation based on morphological preprocessing and 2D-Gabor wavelets. In *Advanced Computing and Intelligent Engineering*; Advances in Intelligent Systems and Computing; Springer: Singapore, 2020; Volume 1082, pp. 411–423. [CrossRef]
60. Ma, W.; Yu, S.; Ma, K.; Wang, J.; Ding, X.; Zheng, Y. Multi-task neural networks with spatial activation for retinal vessel segmentation and artery/vein classification. In *Proceedings of MICCAI*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2019; Volume 11764, pp. 769–778.. [CrossRef]
61. Tan, M.; Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of ICML*; PMLR: Cambridge, MA, USA, 2019; pp. 6105–6114.
62. Lin, T.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In *Proceedings of ICCV*; IEEE: Piscataway, NJ, USA, 2017; pp. 2980–2988.
63. Salehi, S.; Erdogmus, D.; Gholipour, A. Tversky loss function for image segmentation using 3D fully convolutional deep networks. In *Proceedings of MLMI*; Springer: Cham, Switzerland, 2017; pp. 379–387.
64. Szegedy, C.; Vanhoucke, V.; Ioffe, S.; Shlens, J.; Wojna, Z. Rethinking the inception architecture for computer vision. In *Proceedings of CVPR*; IEEE: Piscataway, NJ, USA, 2016; pp. 2818–2826.
65. Lee, C.; Xie, S.; Gallagher, P.; Zhang, Z.; Tu, Z. Deeply-supervised nets. In *Proceedings of AISTATS*; PMLR: Cambridge, MA, USA, 2015; pp. 562–570.
66. Kingma, D.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
67. Yang, Y.; Soatto, S. FDA: Fourier domain adaptation for semantic segmentation. In *Proceedings of CVPR*; IEEE: Piscataway, NJ, USA, 2020; pp. 4085–4095.
68. Schneider, S.; Rusak, E.; Eck, L.; Bringmann, O.; Brendel, W.; Bethge, M. Improving robustness against common corruptions by covariate shift adaptation. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 11539–11551.
69. Galvão, W.N.; Comin, C.H. VessShape: Few-shot 2D blood vessel segmentation by leveraging shape priors from synthetic images. *arXiv* **2025**, arXiv:2510.27646.
70. Seghier, M.L. Image segmentation evaluation with the Dice index: Methodological issues. *Int. J. Imaging Syst. Technol.* **2024**, *34*, e23203. [CrossRef]
71. Müller, D.; Soto-Rey, I.; Kramer, F. Towards a guideline for evaluation metrics in medical image segmentation. *BMC Res. Notes* **2022**, *15*, 210. [CrossRef] [PubMed]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.