

# From Segmentation to Morphometry: A Post-Processed U-Net Framework for Corneal Endothelial Cell Analysis

Rémi Badel<sup>1</sup>, Raphaël Couturier<sup>1</sup>, and Anne-Sophie Gauthier<sup>2,3</sup>

<sup>1</sup>Université Marie et Louis Pasteur, CNRS, institut FEMTO-ST (UMR 6174), Belfort, France

<sup>2</sup>Department of Ophthalmology, CHU Besançon, 3 Bd Fleming, 25000, Besançon, France

<sup>3</sup>Université Marie et Louis Pasteur, INSERM, RIGHT (UMR 1098), Besançon, France

## Abstract

**Problem.** Quantitative assessment of the corneal endothelium relies on morphometric parameters (endothelial cell density, coefficient of variation of cell area, pleomorphism) extracted from specular microscopy images. Manual cell delineation is labor-intensive, observer-dependent, and poorly reproducible, while fully automated segmentation remains challenging because of uneven illumination, low contrast, imaging noise and the limited size of publicly available annotated datasets.

**Aim.** We aim to design a fully automated and reproducible pipeline that produces topologically consistent cell contours suitable not only for accurate pixel-wise segmentation but also for clinically relevant morphometric quantification, and that generalizes to images acquired under different conditions.

**Methods.** We propose a framework that combines a lightweight U-Net adapted to small patches (48×48 pixels) with a dedicated morphological post-processing pipeline (Sauvola adaptive thresholding, largest-component extraction, morphological cleanup, skeletonization with branch pruning). The network is trained on the Gavet dataset of 30 human corneal endothelium specular microscopy images using categorical cross-entropy, the Adam optimizer and early stopping, and applied at inference through an overlapping sliding-window strategy. The framework is evaluated on a held-out Gavet test set and on the external Alizarine dataset (porcine endothelium) which was never seen during training, using the Dice coefficient, the Modified Hausdorff Distance (MHD), Wilcoxon signed-rank tests and Bland–Altman analysis on the three morphometric parameters.

**Results.** On the Gavet test set, segmentation achieved a mean Dice score of  $0.93 \pm 0.01$ , while on the Alizarine dataset, performance remained robust with a mean Dice of 0.81. Morphometric parameters—cell density, size variability (CV), and pleomorphism—were compared to ground truth using Wilcoxon tests and Bland–Altman analysis. Low bias and narrow agreement limits were observed for the Gavet dataset, while the Alizarine dataset showed minor underestimation for density and increased variability for pleomorphism.

**Conclusion.** Coupling a compact U-Net with a domain-specific post-processing pipeline enables accurate and reproducible segmentation as well as reliable morphometric quantification across datasets, including under domain shift. The proposed framework constitutes a practical and lightweight basis for automated corneal endothelial assessment and a step toward integration into routine clinical workflows.

**Keywords:** corneal endothelium; specular microscopy; deep learning; U-Net; image segmentation; morphometric analysis; post-processing; domain shift.

## 1 Introduction

Morphometric analysis of the corneal endothelium plays a crucial role in the assessment of corneal health and in the diagnosis and monitoring of several ocular pathologies, including Fuchs’ endothelial dystrophy. In a healthy cornea, endothelial cells form a regular, predominantly hexagonal mosaic [38] whose structural organization reflects optimal physiological function. Alterations in this organization

constitute early indicators of endothelial dysfunction [32]. Quantitative morphometric parameters such as endothelial cell density (ECD), coefficient of variation of cell area (CV), and the proportion of hexagonal cells (pleomorphism) are therefore widely used in clinical practice to characterize endothelial status. These parameters are typically derived from specular microscopy images. However, their extraction still relies largely on manual or semi-automatic delineation of cell boundaries by trained experts. This process is time-consuming, labor-intensive, and prone to both inter- and intra-observer variability, which limits reproducibility and hinders large-scale or routine clinical deployment. Consequently, the development of robust and fully automated segmentation methods remains a major challenge in ophthalmic image analysis.

Automatic segmentation of the corneal endothelium is inherently difficult due to the characteristics of specular microscopy images. Such images frequently exhibit uneven illumination, low contrast, noise, and acquisition-related artifacts, all of which complicate accurate boundary detection. Although a wide range of approaches has been proposed over the past decades—ranging from classical image processing techniques to machine learning and, more recently, deep learning-based methods—robust and generalizable solutions remain elusive, particularly under heterogeneous imaging conditions.

Convolutional neural networks, and U-Net-based architectures in particular, have demonstrated strong performance for biomedical image segmentation, including corneal endothelial cell delineation. Nevertheless, when applied to small datasets and images of variable quality, raw network predictions often contain local segmentation errors such as cell merging, boundary discontinuities, or spurious contours. While these imperfections may have a limited impact on global segmentation metrics, they can significantly bias downstream morphometric measurements, especially shape-related parameters such as pleomorphism.

Despite a growing body of work on deep-learning-based endothelial segmentation [12, 41, 40, 21, 44, 6, 43, 10], several issues remain only partially addressed: (i) datasets are small and relatively homogeneous, which limits the ability of high-capacity models to generalize across imaging conditions; (ii) most studies focus on global pixel-wise metrics (Dice, IoU, accuracy) but rarely report the sensitivity of morphometric parameters—which are the quantities of clinical interest—to local segmentation errors; (iii) although post-processing is acknowledged as critical to recover usable cell contours, its design and parameters are often only briefly described, which hinders reproducibility; and (iv) very few studies systematically evaluate cross-dataset generalization, which is the situation actually encountered in clinical practice. These observations motivate the present work.

In this paper, we therefore address the following question: *how can we obtain segmentation maps of the corneal endothelium that are not only accurate at the pixel level but also topologically reliable, so that the derived morphometric parameters remain trustworthy under limited training data and across acquisition conditions?* Our answer is a fully automated framework that explicitly couples a lightweight U-Net architecture, tailored to small patches and limited training data, with a dedicated morphological post-processing pipeline whose role is to enforce topological consistency on the predicted contours. The framework is evaluated both on a held-out test set from the training domain (Gavet) and on an external unseen dataset (Alizarine), enabling a joint assessment of segmentation accuracy, robustness to domain shift, and the reliability of the derived morphometric parameters.

The main contributions of this work can be summarized as follows:

- **Lightweight U-Net adapted to small datasets.** We design a reduced U-Net architecture operating on  $48 \times 48$  patches, with fewer encoder/decoder stages, a progressively decreasing number of filters and dropout regularization, which makes training feasible on the Gavet dataset (30 images) without compromising accuracy.
- **Topology-aware post-processing pipeline.** We propose an explicit, fully reproducible post-processing pipeline—combining morphological dilation, Sauvola adaptive thresholding, largest-component extraction, morphological cleanup and skeletonization with branch pruning—that turns raw probability maps into topologically consistent cell contours suitable for morphometric analysis.
- **Evaluation under domain shift.** We assess segmentation and morphometric performance on both the Gavet test set and an external unseen Alizarine dataset, providing a realistic measure of cross-dataset generalization. We complement standard pixel-wise metrics (Dice, MHD) with morphometric agreement metrics (Wilcoxon signed-rank tests, Bland–Altman analysis).

- **Clinically motivated morphometric analysis.** We demonstrate that on the Gavet test set ECD and CV are estimated with low bias and narrow limits of agreement, while the Alizarine dataset reveals which parameters (notably pleomorphism) are most sensitive to segmentation imperfections, providing useful guidance for the clinical interpretation of automated outputs.
- **Reproducibility-oriented description.** We provide a complete description of the data, pre-processing, training, inference (sliding window), post-processing and morphometric computation, designed to allow other groups to reproduce or extend the framework.

The remainder of the paper is organized as follows. Section 2 reviews related work on corneal endothelium segmentation, highlighting the strengths and limitations of existing approaches. Section 3 describes the proposed framework in detail (data, preprocessing, U-Net architecture, training, inference, post-processing, morphometric parameters). Section 4 reports qualitative and quantitative results on the Gavet and Alizarine datasets. Section 5 discusses the limitations of the proposed approach. Section 6 summarizes our findings and outlines future research directions.

## 2 Related work

The process of image segmentation of the corneal endothelium has evolved significantly over the last three decades, with a transition from classical image processing techniques to the most complex and advanced deep learning architectures. In what follows we organize the literature into four families—traditional methods, conventional machine learning, convolutional neural networks, and recent transformer-based or hybrid approaches—and, for each family, we explicitly highlight the main strengths and limitations as well as the resulting research gaps motivating the present work.

### Traditional methods

Initial approaches for the segmentation of endothelial cells in specular microscopy images relied on simple techniques such as thresholding, skeletonization, Gaussian filtering, and morphological operations to delineate cell contours. For instance, Ayala *et al.* [2] and Nadachi [26] used a combination of simple binarisation and morphological filtering. Directional filters have also been proposed by Habrat *et al.* [16] and Mazhoun [23] to improve the visibility of hexagonal cells before applying logic rules. In parallel, Sanchez-Marín [33] proposed a fully automated segmentation approach using Gaussian filtering and contrast enhancement, which offered greater robustness to variations in image quality. However, these early methods remained limited by variations in image quality and required human intervention.

Watershed-based methods have become popular due to their ability to effectively identify closed regions in images, making them particularly useful for segmentation tasks. Several variations have been explored to enhance this approach, such as incorporating markers [42] to guide the segmentation process, using distance maps [1, 4] to refine boundaries, and employing Fourier analysis and applying stochastic methods [36] to account for uncertainties during segmentation. However, watershed-based methods are often associated with oversegmentation, i.e. the capacity to divide the image into an excessive number of segments, thereby reducing the accuracy of segmentation in certain cases. To address this issue, Issam and Kamal [9] proposed combining watershed transformation with active contours (or snakes), which help refine the segmented boundaries by adapting more precisely to the shapes of objects. Unfortunately, these approaches remain sensitive to oversegmentation and noise.

*Advantages.* Traditional methods are computationally cheap, fully interpretable, do not require labelled data and can be deployed without GPU. *Disadvantages.* They are highly sensitive to illumination variations and noise, require careful manual tuning of parameters per image or per dataset, do not enforce any high-level shape prior, and typically need an expert to validate or correct the result, which limits their use in routine clinical practice.

### Conventional machine learning approaches

In the era of machine learning, numerous techniques have been proposed to overcome the limitations of traditional ruled-base methods. Foracchia and Ruggeri [14] applied Bayesian classification, while Poletti *et al.* [30] used support vector machines (SVM) to identify cell regions. In their study, Scarpa

and Ruggeri [35] developed a segmentation approach for corneal endothelial images based on a genetic algorithm. It simulates natural selection to optimize cell contour detection by combining pixel intensity and the regular geometric structure of cells. This innovative strategy has been developed to facilitate the calculation key clinical parameters such as cell density, pleomorphism, and polymegethism, also ensuring a greater proximity with expert annotations. Despite their enhanced flexibility afforded by these methodologies, a substantial degree of manual intervention remains a prerequisite.

*Advantages.* Compared with rule-based methods, classical learning approaches better handle local intensity variations, provide a probabilistic interpretation of cell membership and can incorporate prior shape information through hand-crafted features. *Disadvantages.* They rely on hand-engineered features whose discriminative power is limited under heterogeneous acquisition conditions, often require region-specific post-processing, and still depend on partial human supervision; their generalization to unseen datasets has not been systematically demonstrated.

## Convolutional neural network approaches

Recent advances in deep learning, particularly using convolutional neural networks (CNNs), have led to significant progress in the segmentation of corneal endothelial cells. Two main approaches can be identified in the utilization of CNNs for this type of segmentation:

**Pixel-by-pixel classification** This strategy is generally implemented using a sliding window, where each pixel is analysed independently according to its local neighbourhood. This strategy was employed in the initial research, notably by Katafuchi and Yoshimura [20], who segmented *in vivo* images of the human endothelium, and by Nurzynska [27, 28], who worked on *ex vivo* images acquired by phase contrast microscopy and proposed a variant of the previous method by introducing a third category of pixels corresponding to cell centres, thereby enabling more precise distinction between cells. The two approaches under discussion were both based on the network architecture proposed by Ciresan *et al.* [8] and achieved excellent results in terms of segmentation. However, it should be noted that Katafuchi’s work does not include any clinical validation. Concurrently, Fabijanska [11] proposed a feed-forward neural network with a single hidden layer on the same *ex vivo* dataset.

**Patch-to-patch classification** This approach, especially in the context of corneal endothelial images, involves the division of the image into small regions, known as patches. Then, a convolutional neural network (CNN) is applied to each patch. Unlike the previous method, this allows for a better understanding of the overall spatial context of the image, which improves edge detection and the identification of complex relationships between image structures. The use of encoder-decoder architectures, such as the well-known U-Net network, is commonly used for this task. This approach was employed by Fabijanska [12] to segment specular microscopy images, yielding impressive results highlighting the effectiveness of this approach compared to older methods which fail to capture the overall image context as accurately and efficiently.

Finally, Vigueras *et al.* [41] compared these two approaches for corneal endothelium image segmentation. The U-Net approach, processing the image globally, can be considered a patch-to-patch method, while the SW-net, using a sliding window, is a pixel-to-pixel method. Both methods were evaluated on 50 images acquired with a specular microscope, and their results were compared to manual delineations made by a trained observer.

*Advantages.* CNN-based approaches considerably outperform earlier methods in terms of pixel-wise segmentation accuracy, do not require hand-crafted features, and can be retrained for new acquisition modalities. Encoder-decoder architectures such as U-Net additionally exploit multi-scale context. *Disadvantages.* They typically require sizeable annotated datasets to reach their best performance, are sensitive to domain shifts, and the raw probability maps they produce often contain topological artefacts (broken contours, merged cells) that strongly bias morphometric estimates if not properly post-processed. Several recent works [19, 13, 43] have addressed efficiency or training-data scarcity, but rarely the joint question of segmentation quality *and* morphometric reliability.

Image segmentation often benefits from a post-processing phase to refine the model’s predictions. This essential step improves the quality of predictions by removing noise and correcting local errors using additional algorithms. Although these techniques are effective, post-processing is not explicitly discussed in many studies, which limits reproducibility and comparability across publications.

## Corneal endothelium segmentation: issues with image quality and data scarcity

Segmentation of corneal endothelial cells is a major challenge due to image variability and the small amount of annotated data available. With this in mind, Kucharski and Fabijanska [21] explored the use of generative adversarial networks (GANs) to augment training data, assessing their ability to diversify datasets while preserving the biological plausibility of cell structures. Their approach, applied specifically to the segmentation of corneal endothelium, makes it possible to generate synthetic images without resorting to exhaustive manual annotation. It thus helps to overcome the lack of annotated data and to improve the performance of segmentation models in different contexts.

Recently, Vigueras [40] has used an improved version of the U-Net architecture, called DenseUNet, which combines feedback (allowing predictions to be refined by reusing information from deeper layers to adjust results from higher layers) and non-local attention (which captures long-distance relationships between pixels). This method improves the segmentation of corneal endothelial cells, particularly in complex conditions, such as the presence of guttae or cellular deformations characteristic of pathologies such as Fuchs' dystrophy. In a follow-up work, the same group [39] reformulated the segmentation problem as the regression of signed distance maps, achieving robust performance across all stages of Fuchs' dystrophy.

Images of the corneal endothelium are frequently disrupted by irregular reflections and subject movement, blurring cell edges and complicating segmentation. To overcome these image quality difficulties, the hybrid multi-branch MBT-Net proposes an innovative architecture, combining a convolutional block for local feature extraction with a Transformer for establishing long-range dependencies in the image. Using this approach, Zhang *et al.* [44] achieved significant improvements over conventional methods, particularly on the Alizarine dataset. More general medical-imaging transformer-based architectures, such as Swin-UNet [5] and TransUNet [7], have been proposed for biomedical image segmentation; they typically achieve strong results when trained on large datasets but require substantially more parameters and training data than convolutional U-Nets, and have not been specifically validated on small specular-microscopy datasets such as Gavet.

### Recent advances (2023–2026)

Recent work has focused on three complementary directions: efficiency, robustness to data scarcity, and clinical deployment. Karmakar *et al.* [19] introduced MOBILE-CELLNET, a hybrid deep model combining lightweight CNN encoders with classical post-processing, demonstrating that compact models are sufficient for endothelial segmentation. Al-Fahdawi *et al.* [13] proposed CELLSDEEPNET, a web-based pipeline integrating segmentation and morphometric analysis. To overcome data scarcity, Kucharski and Fabijanska [21] relied on GAN-based synthetic data, while Wang *et al.* [43] showed that warping-based augmentation considerably increases segmentation accuracy in semi-supervised settings. Cassinerio *et al.* [6] and Daniels-Tomaczak *et al.* [10] reported deep-learning analyses of widefield endothelial imaging in clinical contexts (DMEK, Fuchs dystrophy), confirming the practical interest of automated pipelines. A 2026 review [15] and a recent systematic review focused on Fuchs dystrophy [18] jointly emphasize that, despite this progress, two questions remain open: (i) the reliability of morphometric parameters derived from automated segmentation, and (ii) the robustness of these parameters under inter-dataset and inter-device variability. The present work directly addresses these two open issues.

*Advantages of recent transformer/hybrid approaches.* They model long-range dependencies, handle complex pathological cases (guttae, Fuchs dystrophy) and benefit from large-scale pretraining. *Disadvantages.* They are heavier to train and deploy, often require careful regularization on small datasets, and—to the best of our knowledge—have not been systematically benchmarked on the small but well-annotated Gavet dataset together with cross-dataset evaluation on Alizarine.

### Identified research gaps

Building on the above review, we identify the following gaps that motivate the present study:

1. **Reproducibility of post-processing.** Many pipelines mention post-processing only briefly; a detailed and reproducible specification of the operations and parameters is rarely provided.

2. **Topological reliability of contours.** Most works report pixel-wise metrics (Dice, IoU) but do not quantify the impact of local segmentation errors on derived morphometric parameters, which are the actual clinical outputs.
3. **Cross-dataset generalization.** Performance under domain shift (different cameras, species, magnification) is rarely reported, even though it is the practical situation faced by clinicians.
4. **Compact architectures for small datasets.** The interaction between architecture capacity, training-data size and the need for post-processing remains underexplored, especially when only a few dozen annotated images are available.

A more comprehensive review of the approaches considered for corneal endothelium segmentation is available in the article [17], and more recent surveys are provided in [15, 18].

## 3 Methodology

### 3.1 General approach

The proposed approach (Figure 1) is based on a multi-stage processing chain, from image pre-processing to the estimation of morphometric parameters. After a process of resizing and contrast enhancement, the images are divided into patches, which are then used to train a convolutional neural network (CNN). The aim of this CNN is to facilitate cell segmentation. The next step in the processing is to refine the raw segmentation map. Finally, morphometric measures such as cell density (ECD), coefficient of variation (CV) and hexagonality (HEX) are extracted from the final segmented maps. The pipeline can be described as five sequential modules, each with a clearly defined input/output relation:

1. **Preprocessing module** (Sec. 3.2). *Input:* raw  $H \times W$  grayscale specular-microscopy image. *Operations:* normalization to  $[0, 1]$ , Contrast-Limited Adaptive Histogram Equalization (CLAHE) and, at training time only, geometric augmentation (horizontal/vertical flips). *Output:* a contrast-enhanced image of the same size whose dynamic range is more favorable for boundary detection.
2. **Patch extraction.** *Input:* preprocessed image. *Operations:* dense extraction of overlapping  $48 \times 48$  patches (training and inference, see Sec. 3.4) with a stride of 4 pixels at training time, with the corresponding patches in the ground-truth mask used as targets. *Output:* a set of input/output patch pairs.
3. **Modified U-Net** (Sec. 3.3). *Input:* a patch (size  $H_p \times W_p$ ). *Operations:* encoder-decoder forward pass with skip connections; produces a per-pixel two-class soft prediction (background vs. cell boundary). *Output:* a probability patch of the same spatial size.
4. **Sliding-window reconstruction** (Sec. 3.5). *Input:* the set of patch-level predictions. *Operations:* averaging of overlapping predictions, performed line-by-line to bound memory consumption. *Output:* a full-image probability map (raw segmentation map).
5. **Topology-aware post-processing** (Sec. 3.6). *Input:* raw probability map. *Operations:* dilation, Sauvola adaptive thresholding, largest-connected-component extraction, morphological cleanup, skeletonization with branch pruning. *Output:* a thin, topologically consistent binary contour map suitable for morphometric quantification.

The morphometric parameters (ECD, CV, pleomorphism) are then computed from the post-processed contour map (Sec. 3.7). The overall pipeline is summarized in Fig. 1: starting from a Gavet image and its associated mask, the data flow proceeds through preprocessing, the CNN, the raw segmentation map, the post-processing block and finally the morphometric module that produces ECD, CV and HEX. Each block in Fig. 1 corresponds to one of the modules described above, and the morphometric block is connected to the final segmentation map to make explicit the fact that quantitative outputs are derived only after topology-aware refinement.

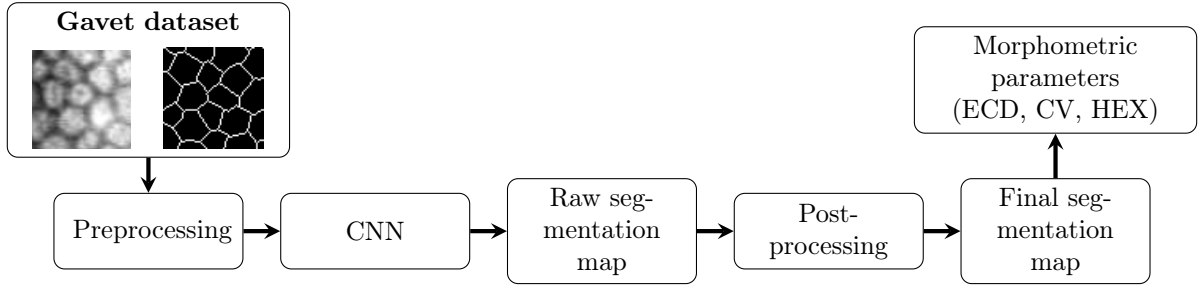


Figure 1: Overview of the proposed pipeline for corneal endothelial cell segmentation and morphometric analysis. Starting from a specular-microscopy image and its annotated mask (Gavet dataset), the pipeline applies (i) preprocessing (normalization, CLAHE), (ii) a modified U-Net producing a raw probability map of cell boundaries, (iii) a topology-aware post-processing block (dilation, Sauvola adaptive thresholding, largest-component extraction, morphological cleanup, skeletonization with branch pruning) producing a final segmentation map, and (iv) a morphometric module that extracts the endothelial cell density (ECD), the coefficient of variation of cell area (CV) and the hexagonality / pleomorphism (HEX) from the final segmentation map.

## 3.2 Data and preprocessing

### Dataset

A dataset of 30 images of the human corneal endothelium, obtained by specular microscopy, was generously provided by Professor Yann Gavet<sup>1</sup>, to whom we express our sincere thanks. The images are 8-bit grayscale with a relatively low resolution of  $152 \times 388$  pixels, but they are completely filled with endothelial cells. In comparison, the Alizarine dataset - frequently referenced in the literature [32] or <http://bioimlab.dei.unipd.it> - provides images at a higher resolution of  $768 \times 576$  pixels, but only a limited region is well focused on the cellular structures.

Each image in the dataset is accompanied by carefully annotated ground truth, providing manually traced cell contours. This combination of high annotation quality and cellular density makes the dataset particularly suitable for the training, validation, and evaluation of image segmentation methods.

**Intra-dataset variability of Gavet.** The 30 Gavet images were acquired on *healthy or near-healthy human corneas*; they cover physiological inter-individual variability in cell density, cell size and image contrast, but they do not include cases of advanced corneal pathology such as advanced Fuchs' dystrophy with extensive guttae or post-surgical endothelial remodelled. As a consequence, the present study evaluates the proposed framework primarily under the regime of healthy or moderately altered endothelium; the extension to severely pathological cases is discussed in Section 5.

While the model is trained and validated on the GAVET dataset, it is subsequently applied to the ALIZARINE dataset to assess generalization on unseen data.

### Preprocessing

The majority of images obtained by specular microscopy suffer from irregular lighting and generally low contrast, making it difficult to segment cell contours accurately. In order to overcome these limitations and optimize the neural network training, several preprocessing steps were implemented. First, the images were normalized to bring pixel intensities into the range of values between 0 and 1, which reduces the model's sensitivity to scale variations between images and allows for faster convergence during training. Next, the images were subjected to a local contrast enhancement process called CLAHE (Contrast-Limited Adaptive Histogram Equalization) (Figure 2) to better highlight cell contours. Using a kernel size of  $24 \times 24$  pixels, approximately the average cell size, this step allows for better highlighting of relevant details, while controlling noise amplification in homogeneous regions. Finally, a data augmentation strategy was adopted to accommodate the limited size of the training set: horizontal and vertical symmetries were applied, quadrupling the number of examples while preserving the integrity of cellular structures. More complex transformations, such as rotations

<sup>1</sup>LGF CNRS 5307 – École Nationale Supérieure des Mines de Saint-Étienne, 158 Cours Fauriel, CS 62362, 42023 SAINT-ÉTIENNE cedex 2

or elastic deformations, were discarded to avoid the introduction of unrealistic artifacts that could affect the quality of the segmentation. In particular, elastic deformations were deliberately excluded because, although they are commonly used in biomedical image segmentation, they can locally distort the hexagonal arrangement and the apparent cell area, which would directly bias the morphometric parameters (ECD, CV, pleomorphism) that constitute the clinical output of our pipeline. We also chose not to apply rotations of arbitrary angle, since the Gavet acquisition protocol produces images with a consistent orientation and rotation augmentation would force the model to learn invariances that are not present in the deployment setting. The resulting augmentation policy preserves cellular morphology exactly while still quadrupling the effective training-set size.

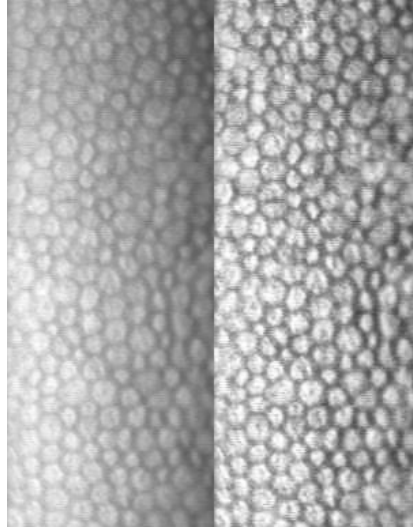


Figure 2: Specular microscopy image ( $388 \times 152$  pixels) before (left) and after (right) contrast enhancement using the CLAHE method.

### 3.3 Model architecture

The architecture used for corneal endothelial image segmentation (Figure 3) is derived from the U-Net network, proposed by Ronneberger *et al.* in 2015 [31], specifically designed for biomedical image segmentation. U-Net is based on a symmetric encoder-decoder structure and uses only convolutional layers. The network’s contraction path reduces image size while enriching representations by extracting increasingly abstract features. The expansion path, on the other hand, progressively reconstructs the image in high resolution. It uses direct connections with the corresponding encoder layers to preserve fine details and improve segmentation accuracy.

In order to take account of the small size of the patches ( $48 \times 48$  pixels), the U-Net architecture was modified by removing two convolution blocks and their associated pooling/upsampling layers. The number of filters in the encoder is progressively doubled from 64 to 256 (mirrored in the decoder), and dropout layers were added to limit overfitting. This simplified version maintains the strengths of U-Net, while being faster to train and better suited to small datasets.

**Detailed modifications with respect to the vanilla U-Net.** For clarity, we summarize the differences between the original U-Net [31] and the architecture used here:

- **Encoder/decoder depth.** The vanilla U-Net uses four downsampling and four upsampling stages. To match the small spatial resolution of our  $48 \times 48$  patches and to avoid trivial bottlenecks (the deepest feature map would otherwise have a resolution of  $3 \times 3$ ), we use only *two* downsampling and two upsampling stages, leading to a feature map of  $12 \times 12$  at the bottom of the U.
- **Number of filters.** The vanilla U-Net doubles the number of filters at each stage starting from 64 (64, 128, 256, 512, 1024). We instead use a much narrower model with progressively (64, 128, 256) filters and a corresponding mirrored decoder. This reduces the parameter count by

more than an order of magnitude and is therefore better suited to the limited size of the training set.

- **Convolution block.** Each convolution block consists of two  $3 \times 3$  convolutions with ReLU activations, separated by a dropout layer with rate 0.1, as illustrated in the inset of Figure 3. The dropout layers act as a regularizer and were not present in the original U-Net.
- **Skip connections.** As in vanilla U-Net, encoder feature maps are concatenated channel-wise with their decoder counterparts before each convolution block of the expansive path; we did *not* add attention gates or feedback non-local attention modules [40], in order to keep the model lightweight on a  $48 \times 48$  patch.
- **Output layer.** A final  $1 \times 1$  convolution followed by a softmax produces a 2-channel probability map (background vs. cell boundary) with the same spatial size as the input patch.

The resulting model contains 2,210,114 trainable parameters ( $\sim 2.2$  M), about  $3\times$  fewer than SwinUNet (6.3 M) and comparable in size to a lightweight TransUNet (1.5 M) [5, 7], which is consistent with the limited size of the training set and the small patch resolution used.

### 3.4 Training and optimization

The network was implemented in Python 3.12 using the PyTorch library. The input images were resized to  $388 \times 152$  pixels and then divided into  $48 \times 48$  pixel patches with a 4 pixel stride, generating 2 322 random patches per image.

The dataset used in this study contained 30 images, divided into three subsets in the following proportions: 64% for training, 16% for validation, and 20% for testing. The neural network was trained using patches of  $48 \times 48$  pixel images, randomly extracted from the training images with a 4 pixel stride, at a rate of 2 322 patches per image. Other choices of size were possible, but this format provided the best compromise between pixel classification accuracy and computational load, although the training time per epoch was longer. Training was carried out over 150 epochs, with a batch size set at 32. The network weights were optimized using the Adam algorithm (the initial learning rate was set to  $10^{-3}$ ), by minimizing a loss function based on categorical cross-entropy. In addition, evaluation metrics such as accuracy, DICE score, AUC, as well as loss, were monitored on the training, validation, and test sets. Unlike the loss function, which directly guides the optimization, these metrics allow the overall performance of the model to be assessed. Finally, an early stopping mechanism was implemented to prevent overfitting: training was stopped, and the model corresponding to the best performance was saved. The training was carried out on an NVIDIA A100-PCIE-40GB GPU, which considerably reduced the learning time.

**Loss function.** Each pixel of a patch is classified into one of two classes,  $c \in \{0, 1\}$  (background and cell boundary). Let  $p_i^{(c)}$  denote the predicted probability that pixel  $i$  belongs to class  $c$ , and  $y_i^{(c)} \in \{0, 1\}$  the corresponding ground-truth indicator (one-hot encoding). The training objective is the categorical (binary) cross-entropy averaged over all pixels of all patches in the mini-batch:

$$\mathcal{L}_{\text{CE}} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=0}^1 y_i^{(c)} \log p_i^{(c)}, \quad (1)$$

where  $N$  is the total number of pixels in the batch. The choice of cross-entropy was driven by two considerations. First, it provides a smooth, well-behaved gradient even when the predicted probability is close to 0 or 1, which is helpful for the small patches considered here. Second, the foreground/background class imbalance, while pronounced (cell-boundary pixels represent on average  $6.4 \pm 1.6\%$  of a Gavet image, range 3.7–8.8%), is consistent across images and could be handled without resorting to specialized losses. We did consider compound losses such as Dice [25] or Focal [22] losses, but on this dataset cross-entropy combined with the post-processing pipeline produced the best balance between training stability and morphometric reliability.

**Hyperparameter rationale and convergence.** We summarize in Table 1 the main hyperparameters and the rationale behind each choice. The patch size of  $48 \times 48$  during training was chosen empirically: smaller patches did not capture enough cells per patch (the average cell diameter is approximately 24 pixels), while larger patches required deeper networks and produced fewer training

samples per image. The stride of 4 pixels both densely covers each image and ensures that during inference any pixel is predicted by multiple overlapping patches, which stabilizes the raw probability map (see Sec. 3.5). The Adam optimizer was preferred over plain SGD because the small batch size and the presence of dropout introduce noisy gradients that Adam handles well; the initial learning rate of  $10^{-3}$  is the canonical default for Adam and was found to be stable across runs. Early stopping was triggered on the validation Dice with a patience of 10 epochs; in practice the best model was reached well before 150 epochs, after which the validation Dice plateaued. No explicit learning-rate schedule (cosine, step) was found necessary on top of Adam’s adaptive behaviour.

Table 1: Training hyperparameters and rationale.

| Hyperparameter          | Value                 | Rationale                                       |
|-------------------------|-----------------------|-------------------------------------------------|
| Patch size (training)   | $48 \times 48$        | $\sim 2$ cell diameters; trade-off context/cost |
| Patch stride            | 4 pixels              | Dense coverage; $\sim 2,322$ patches/image      |
| Train/Val/Test split    | 64 / 16 / 20 %        | Standard small-data split (30 images)           |
| Batch size              | 32                    | Largest size that fits comfortably in GPU       |
| Optimizer               | Adam                  | Robust default for noisy gradients              |
| Initial learning rate   | $10^{-3}$             | Adam canonical default                          |
| Loss function           | Cross-entropy (Eq. 1) | Smooth gradients; mild class imbalance          |
| Max epochs              | 150                   | Upper bound; early stopping below               |
| Early stopping patience | 10 epochs             | Stops when val. Dice plateaus                   |
| Dropout rate            | 0.1                   | Mild regularization on small dataset            |
| Hardware                | NVIDIA A100 40GB      | $\sim 10$ min total training time               |

### 3.5 Prediction

A sliding window approach is used for inference (prediction on a new image). The image is initially divided into small fixed patches of size  $48 \times 48$  pixels with a stride of 4 pixels. Each patch is then processed independently by the model, which predicts a probability map for each pixel. The complete image is then reconstructed by merging the averaged predictions in the overlap areas, which allows for smoothing the transitions between patches. In order to limit memory usage while maintaining accurate reconstruction, fusion is performed line by line using a stride of 1 pixel in the horizontal direction and 2 pixels in the vertical direction. Unlike an attempt at global reconstruction, which would require all the predictions to be loaded into memory - which can be more expensive for large images - this strategy enables the data to be processed sequentially, line by line, considerably reducing the memory load while preserving optimal prediction quality.

**Why a sliding window is needed.** Two reasons make a sliding-window strategy necessary at inference time. First, the network was trained on  $48 \times 48$  patches and operates with a fixed input size; feeding a full image of size up to  $768 \times 576$  pixels (Alizarine) directly into the network would either require retraining at the new resolution or padding the network with cropping, both of which would degrade performance. Second, even if the network were fully convolutional, storing the simultaneous predictions of overlapping patches over an entire high-resolution image quickly exceeds the GPU memory budget for batched inference. The sliding window naturally factors the computation into manageable chunks while *increasing* robustness, since any pixel of the final raw map is the average of multiple independent predictions.

**Choice of window size and stride.** The window size at inference is kept identical to the training patch size ( $48 \times 48$  pixels) so that the input distribution at inference matches the one seen during training. The horizontal stride of 1 pixel and vertical stride of 2 pixels were chosen as a trade-off between the redundancy of the predictions (which improves robustness) and the total inference cost: a horizontal stride of 1 ensures very fine averaging in the direction along which cells are most elongated in the Gavet images, while a slightly larger vertical stride controls the number of overlapping patches.

**Handling overlapping regions.** Each pixel  $(x, y)$  in the output probability map  $P$  accumulates the predictions of all patches  $P_k$  in which it appears. We maintain two buffers: a probability accumulator  $S(x, y)$  and a count map  $C(x, y)$ . For every patch  $k$  centered at position  $(x_k, y_k)$  and every pixel  $(x, y)$  inside the patch:

$$S(x, y) \leftarrow S(x, y) + P_k(x - x_k, y - y_k), \quad C(x, y) \leftarrow C(x, y) + 1. \quad (2)$$

The final probability map is obtained as  $P(x, y) = S(x, y)/C(x, y)$ . To bound memory usage, this accumulation is carried out line-by-line: only the rows of  $S$  and  $C$  currently overlapped by active patches are kept in memory; once a row is no longer reached by any new patch, it is finalized and freed. This streaming reconstruction keeps the memory footprint proportional to the patch height rather than to the image height, which is critical for the Alizarine dataset where images are nearly four times higher than the Gavet ones.

**Differences with training.** The training procedure samples patches randomly from each image (with their associated labels) and uses dropout, while inference (i) deterministically scans patches on a regular grid covering the full image, (ii) disables dropout, and (iii) averages overlapping predictions. No re-normalization of the input intensities is performed at inference: the same normalization  $\rightarrow$  CLAHE pipeline used at training is applied to each test image before patch extraction, so that the input distribution remains consistent.

### 3.6 Post-processing stage

The raw segmentation maps of the cell contours produced by the model highlight the model’s ability to detect the main structures ; however, they still show the need for a post-processing stage to refine the segments, remove artifacts, and obtain a representation that can be used for morphometric analysis. The steps in this post-processing (Figure 4) pipeline are as follows:

1. the **raw segmentation map** ((a) in Figure 5) is the initial output of the model. Although it reveals the overall shape of the cells, it often contains noise and blurred contours ;
2. **morphological dilation** is then applied to enlarge the positive regions of the binary image (i.e. areas containing the detected cell contours) and connect the disjoint areas, thereby improving the continuity of the structures.

The **adaptive thresholding** method proposed by Sauvola et al. [34, 37], is used to binarise the image, i.e. to separate pixels corresponding to cell contours from those belonging to the background. Unlike global thresholding - which applies a single threshold to the entire image, generally based on a global average or statistic but ineffective at properly separating the background and objects in the whole area - Sauvola thresholding relies on a local threshold, calculated specifically for each pixel taking into account its immediate neighborhood. More precisely, for each pixel  $(x, y)$ , a threshold  $T(x, y)$  is defined as a function of the local mean  $m(x, y)$  and the local standard deviation  $\sigma(x, y)$  of the intensities in a window  $w \times w$  centered on that pixel according to the equation (3):

$$T(x, y) = m(x, y) \left[ 1 + k \left( \frac{\sigma(x, y)}{\sigma_{max}} - 1 \right) \right] \quad (3)$$

where  $\sigma_{max}$  is the maximum expected value for the standard deviation (typically 128 for an 8-bit image) and  $k$  is a sensitivity parameter typically between 0.2 and 0.5.

This approach is particularly robust to variations in illumination: in high-contrast areas, the threshold tends towards the local mean, while in low-contrast areas it is reduced, making it easier to eliminate background elements ;

3. once the image has been binarised, filtering is carried out to preserve only the **largest related component**, eliminating small and irrelevant noise regions.
4. **morphological cleaning operations**, such as opening, closing or removing small objects, are then applied to refine the contours and simplify the overall structure ;
5. finally, **skeletonisation** is performed to reduce the contours to a minimum linear representation, followed by **pruning** [29] to remove superfluous secondary branches. The final segmentation map ((b) in Figure 5) then contains a clean and refined structure that is ideally suited to determining morphometric parameters after the fact.

In the Figure 5, the raw segmentation map (a) still contains noise, whereas the final segmentation map (b) shows the improved result after post-processing.

**Algorithmic specification.** For reproducibility, Algorithm 1 provides the complete post-processing pipeline. We denote by  $P \in [0, 1]^{H \times W}$  the raw probability map produced by the network and by  $B \in \{0, 1\}^{H \times W}$  the final binary skeletonized contour map. The structuring elements  $\mathcal{D}_r$  and  $\mathcal{O}_r$  are disks of radius  $r$ . The Sauvola threshold (Eq. 3) is applied with a window of size  $w \times w$  and a sensitivity parameter  $k$ . We set  $w = 17$ ,  $k = 0.20$  for Gavet (small images, regular contrast) and  $w = 37$ ,  $k = 1.10$  for Alizarine (larger images with an unfocused periphery that requires a much more aggressive threshold to be suppressed); these values were the only dataset-specific parameters that had to be re-tuned. Skeletonization uses Zhang–Suen thinning followed by DSE-based pruning [29] with a pruning length threshold of 150 pixels.

---

**Algorithm 1** Topology-aware post-processing.

---

**Require:** Raw probability map  $P \in [0, 1]^{H \times W}$

Parameters: dilation radius  $r_d$ , Sauvola window  $w$ , sensitivity  $k$ , pruning length  $\ell$

**Ensure:** Binary skeletonized contour map  $B$

```

1:  $M \leftarrow P \oplus \mathcal{D}_{r_d}$  ▷ morphological dilation
2: for each pixel  $(x, y)$  do ▷ Sauvola adaptive thresholding
3:   Compute local mean  $m(x, y)$  and standard deviation  $\sigma(x, y)$  in window  $w \times w$ 
4:    $T(x, y) \leftarrow m(x, y)[1 + k(\sigma(x, y)/\sigma_{\max} - 1)]$ 
5:    $B_1(x, y) \leftarrow \mathbf{1}[M(x, y) > T(x, y)]$ 
6: end for
7:  $C \leftarrow \text{LargestConnectedComponent}(B_1)$ 
8:  $C \leftarrow C \circ \mathcal{O}_1$  ▷ morphological opening
9:  $C \leftarrow \text{RemoveSmallObjects}(C, \text{min size} = 8 \text{ px})$ 
10:  $S \leftarrow \text{Skeletonize}(C)$  ▷ Zhang–Suen
11:  $B \leftarrow \text{PruneBranches}(S, \text{length} < \ell)$  ▷ DSE-based pruning
12: return  $B$ 

```

---

### 3.7 Cell morphometric parameters

Morphometric analysis - *morpho* for shape/structure and *metric* for quantitative measurement - of the corneal endothelium is based on the extraction of numerical parameters enabling the cellular organisation to be characterised and the state of this essential layer of the cornea to be assessed. In general, three morphometric parameters are studied:

- **Endothelial cell density (ECD)** is defined as the number of cells per unit area and is considered a fundamental marker of corneal health. It is particularly high at birth, reaching more than 6 000 cells/mm<sup>2</sup> during the first year of life, then gradually decreases by around 0.6% per year from the age of 15, due to the cells' inability to regenerate, to reach around 2 300 cells/mm<sup>2</sup> at the age of 85. In a healthy adult, the average cell density varies between 2 000 and 3 500 cells/mm<sup>2</sup>. Reduced cell density can signal an alteration in the function of the corneal endothelium, often linked to cell ageing, corneal pathologies such as Fuchs' dystrophy, or cell loss following trauma or surgery. This reduction in cell density can lead to a loss of corneal transparency, which can seriously affect vision [3, 24].
- In healthy corneal endothelium, the majority of cells adopt a regular hexagonal shape, reflecting optimal cell organisation and normal cell function. In circumstances where the corneal endothelium is subjected to stress (i.e. in the presence of pathology, surgery, or age), or where there is cell loss, the remaining cells undergo expansion or contraction in order to fill the resulting empty spaces. This phenomenon, referring to an abnormal variation in cell size, is known as **polymegethism** and is defined by the **coefficient of variation of cell size (CV)** value determined with equation (4):

$$\text{CV} = \frac{\text{Std}_{\text{cell area}}}{\text{Mean}_{\text{cell area}}} \times 100 \quad (4)$$

where Std is the standard deviation, which measures the variability of cell sizes in the image, and Mean is the average area of the cells in the image. This parameter then becomes an early marker

of endothelial dysfunction, often detectable even before a significant decrease in cell density (ECD), thus:

- low CV ( $\leq 30\%$ ): the cells are of uniform size, a sign of healthy endothelium ;
  - moderate CV (30-40%): slight variability in size, which may indicate the beginning of cellular stress ;
  - high CV ( $> 40\%$ ): high size variability, indicating altered cell function, as in Fuchs’ dystrophy or after surgery.
- **Pleomorphism** of the corneal endothelium refers to variations in the shape of the endothelial cells. In a healthy subject, the cells have a hexagonal shape, which ensures that the endothelium functions optimally. But in the event of cell loss (due to disease, trauma or surgery, age), the remaining cells expand to fill the remaining gaps, which inevitably changes their shape. Pleomorphism is assessed by the **hexagonality coefficient (HEX)**, which measures the percentage of hexagonal cells in the endothelium according to the equation 5:

$$\text{HEX} = \frac{N_{\text{hexagonal cells}}}{N_{\text{total cells}}} \times 100 \quad (5)$$

where  $N_{\text{hexagonal cells}}$  is the number of hexagonal cells and  $N_{\text{total cells}}$  is the total number of cells. In a normal state, around 60% of cells are hexagonal. Values below 60% may indicate pathologies and increased impairment of endothelial function affecting corneal transparency and vision.

**Computational pipeline for the morphometric parameters.** The three parameters above are extracted from the post-processed binary contour map  $B$  produced by Algorithm 1. Algorithm 2 formalizes their computation. Connected-component labeling on the inverted contour map yields the individual cell regions  $\{R_1, \dots, R_K\}$ . Cells touching the image border are discarded to avoid partial-cell bias. From each remaining cell  $R_i$  we extract: its area  $a_i$  (in pixels<sup>2</sup>, then converted into mm<sup>2</sup> using the calibrated pixel size of the device), perimeter  $p_i$ , convex-hull area  $h_i$ , and the lengths of the major and minor axes  $L_i$  and  $\ell_i$  of its best-fit ellipse. From these we compute the three shape descriptors used in our regularity score:

$$\text{Circularity}(R_i) = \frac{4\pi a_i}{p_i^2}, \quad (6)$$

$$\text{Solidity}(R_i) = \frac{a_i}{h_i}, \quad (7)$$

$$\text{Aspect ratio}(R_i) = \frac{L_i}{\ell_i}. \quad (8)$$

The regularity score is then

$$\rho_i = \frac{1}{3} \left( \text{Circularity}(R_i) + \text{Solidity}(R_i) + \frac{1}{\text{Aspect ratio}(R_i)} \right), \quad (9)$$

and a cell is declared morphologically regular if  $\rho_i \geq \rho_{\text{th}}$ , with  $\rho_{\text{th}} = 0.85$  in our experiments. Pleomorphism (HEX) is then the fraction of regular cells. The full procedure runs in linear time in the number of cells; for a typical Gavet image with  $\sim 200$ – $300$  cells, the wall-clock time is less than 1 second on a single CPU core. The procedure relies only on standard image-processing primitives (`scikit-image`) and does not require GPU acceleration.

## 4 Results

To evaluate the performance of our cell segmentation model, we followed a comprehensive pipeline comprising several stages. During preprocessing, the raw images were normalized and filtered to reduce noise and enhance cell contrast. The model then performed segmentation, producing raw maps in which each pixel was assigned either to the background or to a cell boundary.

These raw segmentation maps, however, contained numerous artifacts and noise, particularly around cell boundaries. Post-processing was therefore necessary to refine cell contours, remove small

---

**Algorithm 2** Computation of the morphometric parameters (ECD, CV, HEX).

---

**Require:** Post-processed contour map  $B$ , pixel size  $s$  (mm/pixel), regularity threshold  $\rho_{\text{th}}$

**Ensure:** ECD (cells/mm<sup>2</sup>), CV (%), HEX (%)

```
1:  $\bar{B} \leftarrow 1 - B$  ▷ invert: cells are now foreground
2:  $\{R_1, \dots, R_K\} \leftarrow \text{ConnectedComponents}(\bar{B})$ 
3: Discard cells touching the image border
4: Restrict  $\{R_i\}$  to the region of interest defined by the ground-truth mask
5: for each remaining cell  $R_i$  do
6:    $a_i \leftarrow \text{area}; p_i \leftarrow \text{perimeter}; h_i \leftarrow \text{convex-hull area}$ 
7:    $(L_i, \ell_i) \leftarrow \text{major/minor axis of the best-fit ellipse}$ 
8:   Compute  $\text{Circularity}_i$ ,  $\text{Solidity}_i$ ,  $\text{Aspect ratio}_i$ 
9:    $\rho_i \leftarrow \frac{1}{3}(\text{Circ}_i + \text{Sol}_i + 1/\text{AR}_i)$ 
10: end for
11:  $A_{\text{ROI}} \leftarrow \text{physical ROI area in mm}^2 (= \text{pixel count} \cdot s^2)$ 
12:  $\text{ECD} \leftarrow |\{R_i\}|/A_{\text{ROI}}$ 
13:  $\text{CV} \leftarrow 100 \cdot \text{Std}(a_i)/\text{Mean}(a_i)$ 
14:  $\text{HEX} \leftarrow 100 \cdot |\{i : \rho_i \geq \rho_{\text{th}}\}|/|\{R_i\}|$ 
15: return (ECD, CV, HEX)
```

---

spurious regions, and improve the spatial coherence of segmented objects. This step is crucial for extracting reliable morphometric parameters. During metric extraction, quantitative measures such as the Dice coefficient and Modified Hausdorff Distance (MHD) were computed to assess performance.

The proposed method was evaluated on both an internal test set (the Gavet test set, a separate subset of the training dataset containing images not used for parameter optimization) and an external unseen dataset (the Alizarine dataset, consisting of monochromatic images of porcine corneal endothelium—structurally similar to human endothelium—with a resolution of  $768 \times 576$  pixels and a smaller region of well-focused cells). This evaluation framework enabled assessment of both segmentation performance and the reliability of the derived morphometric parameters under domain shift.

## 4.1 Qualitative results

Figures 6 and 7 show, for each dataset, the entire processing pipeline applied to the images, from the initial prediction generated by the network to the final visualisation of the errors in superimposition. The different steps corresponding to these intermediate outputs have been detailed in the Methodology section.

The first step is to display the raw image restricted to a region of interest (ROI). This area is automatically defined from the reference mask associated with the image, then extracted to retain only clearly observable cells. This procedure allows the evaluation to focus on a biologically relevant space, excluding blurred edges, partially visible cells, and out-of-focus areas. The ROI has two main objectives:

- to prevent edge artefacts from introducing bias into the segmentation metrics ;
- to ensure the comparability of morphometric measurements from one image to another.

The network inference then produces a probability map assigning each pixel a value linked to the presence of cell boundaries. Although the overall structure of the cellular network is generally well captured, some imperfections remain : blurred contours in low-contrast areas, merging of adjacent cells (undersegmentation), appearance of false boundaries (oversegmentation), and variability in results due to differences in resolution and contrast between the Gavet and Alizarine datasets.

To make these predictions usable, a post-processing pipeline is applied to produce a consistent, continuous map that is compatible with a reliable estimate of morphometric properties. This phase confirms that, while the network provides a high-performance basis, additional algorithmic correction remains essential before use in a clinical context. It should be noted that certain parameters of Sauvola’s adaptive thresholding were adjusted between the Gavet and Alizarine datasets in order to adapt to the differences in visual characteristics between the two datasets.

Finally, the superimposition of errors allows for a qualitative visualisation of the discrepancies between the ground truth and the prediction. The areas appear:

- in red when annotated cells are not detected (under-segmentation);
- in blue when a cell is artificially created by the prediction (over-segmentation);
- in white when the correspondence between the two masks is correct.

Despite the use of an ROI and a cropping protocol designed to limit edge artefacts, the visualisations indicate that systematic errors persist in the peripheral areas of the Gavet dataset, where under-segmentation remains particularly frequent. In contrast, this issue is not observed in the Alizarine dataset, where under-segmentation errors predominantly occur in the central regions of the images rather than at the periphery.

This qualitative analysis thus highlights the strengths and limitations of the model: overall consistent segmentation within central regions, but recurring weaknesses in peripheral areas and low-contrast zones. To assess more precisely the impact of these observations on overall performance, we now propose a quantitative analysis based on the calculation of the Dice coefficient within the defined ROIs. This step allows us to objectify the trends identified visually and to compare the performance between the two datasets, Gavet and Alizarine, in order to measure the robustness of the model in the face of data changes.

### Qualitative error analysis: Gavet vs. Alizarine

The drop in Dice coefficient from  $0.93 \pm 0.01$  on Gavet to  $\sim 0.81$  on Alizarine deserves a dedicated qualitative analysis, since both datasets contain monochromatic specular-microscopy-like images of corneal endothelium with hexagonal cell mosaics. By visually inspecting the failure cases (Figure 7, columns (d) and (e)), we identified four sources of error that jointly explain this performance gap:

1. **Domain shift in resolution and cell scale.** The Gavet images are  $152 \times 388$  pixels with cell diameters of approximately 24 pixels, whereas Alizarine images are  $768 \times 576$  pixels with larger cells (typically  $\sim 30$ – $40$  pixels). Although the post-processing parameters were re-tuned for Alizarine, the receptive field of the network—which was learned on Gavet—is best matched to the smaller Gavet cells, leading to occasional cell merging in regions of dense Alizarine cells.
2. **Tissue and stain difference.** Gavet was acquired *in vivo* on human corneas with specular microscopy, while Alizarine consists of porcine corneas stained with alizarine red and imaged in transmission. The cell-boundary appearance therefore differs (light-on-dark in Gavet vs. stain-on-light background in Alizarine), which directly affects the contrast distribution and creates a covariate shift that cannot be entirely compensated for by CLAHE.
3. **Image noise and out-of-focus periphery.** Alizarine images contain large regions where cells are out of focus or partially obscured by stain artefacts. These regions tend to produce broken boundaries in the raw probability map; the post-processing pipeline corrects most of them but cannot recreate cells whose evidence is absent from the input.
4. **Annotation conventions.** Manual annotations on Alizarine sometimes include cells whose contours are not fully visible (slightly out-of-focus cells), while our pipeline conservatively excludes them; this systematic bias contributes to the underestimation of cell count discussed below.

The combination of these four factors—and not a failure of the network or post-processing per se—explains the reduced Dice on Alizarine. The fact that the post-processing was sufficient to maintain robust morphometric agreement (low ECD bias, see Sec. 4.2.8) under such a substantial domain shift is, in our view, a positive result.

### Why does the network undercount cells in Alizarine?

The Wilcoxon test reports a statistically significant negative bias on cell count for Alizarine (Table 8), and Table 9 confirms a relative ECD error of  $-2.94\%$  on average. This systematic undercount has two complementary explanations:

- *Conservative ROI definition.* The ROI is defined from the well-focused region of each image; cells partially intersecting this ROI are excluded by design, while the manual annotators sometimes labelled them.
- *Cell merging in low-contrast areas.* When two adjacent cells are separated by a faint boundary, the network occasionally produces a single connected region after thresholding, which counts as one cell instead of two. This effect is amplified by the larger cell size on Alizarine (the network was trained on smaller Gavet cells).

Both effects act in the same direction (slight undercount) but their magnitude—a relative ECD error of  $\sim 3\%$ —is well within the inter-observer variability reported in the literature for manual cell counting [24, 3], which we therefore consider clinically acceptable.

## Representative success and failure cases

To complement the dataset-level analysis above, Figure 8 contrasts representative *success* and *failure* cases of the complete pipeline, selected as the best- and worst-Dice images of each dataset. The top row shows typical successful segmentations: on well-focused, good-contrast images of both Gavet (Dice 0.95) and Alizarine (Dice 0.85) the predicted mosaic closely matches the ground truth, the overlay being almost entirely white (correct) with only sparse red/blue disagreements, and the recovered honeycomb structure is topologically consistent. The bottom row shows the most challenging cases of each dataset. Even the worst Gavet image (Dice 0.91) remains accurate, its errors being confined to a thin band at the image periphery where cells are partially cut off. The worst Alizarine image (Dice 0.79) is more instructive: under-segmentation (red) appears where adjacent cells merge in low-contrast or out-of-focus areas, while over-segmentation (blue) appears as spurious boundaries in dense regions and near stain artefacts. These failure modes are consistent with the four error sources identified above (domain shift in cell scale, tissue/stain difference, out-of-focus periphery and annotation conventions) and with the limitations discussed in Sec. 5: the post-processing pipeline enforces topological consistency but cannot reconstruct cell boundaries whose evidence is absent from the input image. Importantly, these failures remain localised and have a limited effect on the aggregate morphometric estimates, as quantified by the Bland–Altman analysis (Sec. 4.2.8).

## 4.2 Quantitative results

Following the segmentation performed by the **U-Net** model, the accuracy of the detected cell boundaries was assessed by comparison with manually annotated contours (ground truth, denoted  $G$ ). The objective is to measure the ability of the model to reproduce true cellular borders, which is essential for subsequent morphometric analyses (cell density, cell area, hexagonality). Two complementary metrics were employed:

**DICE Coefficient** The DICE coefficient quantifies the global overlap between the predicted contours ( $E$ ) and the ground truth ( $G$ ). It is defined as:

$$\text{DICE} = \frac{2|E \cap G|}{|E| + |G|} \quad (10)$$

A DICE value close to 1 indicates a high degree of overlap and low global segmentation error. This metric is particularly relevant to assess the overall agreement between prediction and reference boundaries.

**Modified Hausdorff Distance (MHD)** The modified Hausdorff distance measures the maximum localisation error between the predicted and reference contour sets:

$$\text{MHD}(E, G) = \max \left\{ \sup_{x \in E} \inf_{y \in G} d(x, y), \sup_{y \in G} \inf_{x \in E} d(x, y) \right\} \quad (11)$$

where  $d(x, y)$  is the Euclidean distance between points  $x$  and  $y$ . Contrary to DICE, MHD is highly sensitive to local inaccuracies and highlights small segmentation defects (missing or spurious cell borders). In summary, DICE evaluates global consistency, while MHD assesses local geometric precision; both are therefore required for a comprehensive evaluation.

To ensure fair and meaningful metric computation, the following preprocessing steps were applied:

- contour dilation to compensate for discretisation-related localisation shifts ;
- extract the region of interest (ROI) from the ground truth in order to limit the analysis to areas that have actually been annotated ;
- distribution analysis via boxplots to visualise dispersion and detect outliers.

These steps ensure that the metrics accurately reflect the model’s performance and not the limitations of the measurement protocol.

#### 4.2.1 Ablation study

To isolate the contribution of each main component of our framework, we compare three variants on the Gavet test set:

- **Base U-Net:** a vanilla U-Net [31] (4 down/up stages,  $64 \rightarrow 1024$  filters) trained on the same patches with the same loss, and without any post-processing.
- **Base U-Net + post-processing:** the same vanilla U-Net followed by the post-processing pipeline of Algorithm 1.
- **Modified U-Net + post-processing** (this work): the lightweight U-Net described in Sec. 3.3 followed by the same post-processing.

Table 2 reports the mean Dice on Gavet for each variant, all evaluated with the same protocol as the main results. The vanilla U-Net alone yields a low Dice: its raw output contains topologically inconsistent contours (broken cells, double boundaries) that the overlap metric heavily penalises. Adding the topology-aware post-processing roughly doubles the Dice ( $0.318 \rightarrow 0.675$ ), confirming the importance of the refinement stage. Replacing the vanilla backbone by the lightweight U-Net of Sec. 3.3 brings a further large improvement (to 0.930) while using an order of magnitude fewer parameters (2.21 M vs 31.4 M), thanks to its better fit to the small training set and small patches.

Table 2: Ablation study on the Gavet dataset (30 images), evaluated with the same protocol as the main results: sliding-window inference, topology-aware post-processing (Sauvola  $w=17$ ,  $k=0.2$ ) and Dice computed with a 3-pixel tolerance. Dice is mean  $\pm$  SD over the 30 images. The Modified U-Net + post-processing row reproduces the headline Gavet Dice of Table 4.

| Variant                                        | Params        | Dice                                |
|------------------------------------------------|---------------|-------------------------------------|
| Base U-Net (no post-processing)                | 31.4 M        | $0.318 \pm 0.175$                   |
| Base U-Net + post-processing                   | 31.4 M        | $0.675 \pm 0.080$                   |
| <b>Modified U-Net + post-processing (ours)</b> | <b>2.21 M</b> | <b><math>0.930 \pm 0.013</math></b> |

#### 4.2.2 Per-component ablation of the post-processing pipeline

While the previous ablation isolates the contribution of the network backbone and of the post-processing block taken *as a whole*, we now quantify the individual contribution of each component of the topology-aware post-processing pipeline (Algorithm 1). Starting from the complete pipeline, we deactivate one component at a time (leave-one-out) and re-evaluate the Dice coefficient on the Gavet test set and on the external Alizarine dataset, using exactly the same protocol as for the main results. The skeletonization step is the only component that cannot be removed in isolation, since the morphometric module (Algorithm 2) requires a one-pixel-wide contour map as input; it is therefore reported as a mandatory stage rather than ablated. For the Sauvola adaptive thresholding we report the effect of *replacing* it by a global Otsu threshold rather than removing the binarisation altogether, which would be ill-defined.

Table 3 reports the Dice obtained for each variant. The analysis reveals a clear hierarchy among the components. The morphological dilation of the raw probability map is by far the most critical step: removing it collapses the Dice to essentially zero on both datasets ( $0.930 \rightarrow 0.000$  on Gavet,

0.819  $\rightarrow$  0.004 on Alizarine), because without it the thin and partially broken boundary responses are not connected into a single mosaic, so the subsequent largest-component extraction retains only a small fragment. The Sauvola adaptive thresholding is the second most important component: replacing it by a global Otsu threshold reduces the Dice by about 0.04 on both datasets (0.930  $\rightarrow$  0.892 and 0.819  $\rightarrow$  0.779), confirming that a locally adaptive threshold is needed to cope with the uneven illumination of specular microscopy. Branch pruning yields a small improvement on Gavet (+0.008) and is essentially neutral on Alizarine (−0.001), its role being to remove spurious skeleton branches. Finally, the largest-component extraction and the morphological cleanup have a marginal effect on the global Dice (within one standard deviation) once the map has been dilated and adaptively thresholded; their main role is to remove residual fragments and double boundaries that, although nearly invisible to the overlap-based Dice, improve the topological consistency of the contour map on which the downstream morphometric quantification relies. In particular, although removing the morphological cleanup marginally increases the Dice (+0.002 on Gavet, +0.003 on Alizarine), this difference is well within one standard deviation and is therefore not significant; we deliberately keep the cleanup because the small fragments it eliminates would otherwise be counted as spurious micro-cells by the morphometric module, biasing the endothelial cell density and the coefficient of variation. The full pipeline is therefore the best *morphometric* configuration, even if, by a negligible margin, it is not the single best variant in terms of raw boundary overlap.

Table 3: Per-component (leave-one-out) ablation of the post-processing pipeline of Algorithm 1, evaluated on the Gavet test set (30 images) and on the external Alizarine dataset with the same protocol as the main results (Dice, mean  $\pm$  SD where available). Each row deactivates exactly one component of the full pipeline; “Sauvola  $\rightarrow$  Otsu” replaces the adaptive threshold by a global one. Skeletonization is mandatory (required by the morphometric module, Algorithm 2) and is therefore not ablated. The first row (full pipeline) closely matches the headline numbers of Table 4.

| Pipeline variant                                           | Dice (Gavet)                        | Dice (Alizarine)                    |
|------------------------------------------------------------|-------------------------------------|-------------------------------------|
| Full pipeline (ours)                                       | <b>0.930 <math>\pm</math> 0.013</b> | <b>0.819 <math>\pm</math> 0.017</b> |
| w/o morphological dilation                                 | 0.000 $\pm$ 0.000                   | 0.004 $\pm$ 0.008                   |
| Sauvola $\rightarrow$ global Otsu threshold                | 0.892 $\pm$ 0.017                   | 0.779 $\pm$ 0.024                   |
| w/o largest-component extraction                           | 0.928 $\pm$ 0.014                   | 0.819 $\pm$ 0.017                   |
| w/o morphological cleanup (opening + small-object removal) | 0.932 $\pm$ 0.013                   | 0.822 $\pm$ 0.018                   |
| w/o branch pruning                                         | 0.922 $\pm$ 0.015                   | 0.820 $\pm$ 0.017                   |

### 4.2.3 Comparison with transformer-based architectures

We compare our approach to two state-of-the-art transformer-based segmentation networks designed for medical images: **Swin-UNet** [5] and **TransUNet** [7]. Both networks were retrained from scratch on the Gavet training set with the same data augmentation and a comparable training budget. Table 4 summarizes the comparison in terms of Dice on the Gavet test set, Dice on the Alizarine dataset, number of parameters and inference time per image. The proposed method achieves a Dice comparable to or higher than both transformers on Gavet with about  $3\times$  fewer parameters than Swin-UNet and a parameter count comparable to the lightweight TransUNet; on the unseen Alizarine dataset it remains competitive, suggesting that for the small dataset sizes typical of this application a lightweight CNN is at least as effective as a transformer-based architecture. For a strictly fair comparison, the raw probability maps produced by Swin-UNet and TransUNet are fed through the *exact same* post-processing pipeline as our network (Sec. 3.6, Algorithm 1), including, on Alizarine, the more aggressive Sauvola thresholding and the ROI-restricted refinement. This post-processing stage is itself a strong contributor to the final accuracy, so the transformer scores reported in Table 4 already benefit from it; in particular, their slightly higher Dice on Alizarine should be read as the performance of a Gavet-trained transformer *combined with our post-processing*, not of the network alone.

### 4.2.4 Computational cost and inference time

The proposed pipeline was designed to be lightweight and clinically deployable. We report in Table 5 the inference time per image, measured separately for (i) the U-Net forward pass over all overlapping

Table 4: Comparison with transformer-based segmentation networks. Swin-UNet (MONAI Swin-UNETR with input padding to  $64 \times 64$ ) and a lightweight TransUNet were retrained from scratch on the Gavet training set with the same patch sampling and data augmentation, and their raw predictions were post-processed with the *same* pipeline as our method (Sec. 3.6), so the transformer scores already include the contribution of our post-processing. Inference time is per full image, sliding window with the paper’s stride (stride 1 pixel horizontally, 2 pixels vertically, see Sec. 3.5), single NVIDIA A100 40GB.

| Method                             | Dice (Gavet)                        | Dice (Alizarine)  | Params        | Inf. time/image |
|------------------------------------|-------------------------------------|-------------------|---------------|-----------------|
| Swin-UNet [5]                      | $0.800 \pm 0.088$                   | <b>0.842</b>      | 6.30 M        | 14.5 s          |
| TransUNet [7]                      | $0.818 \pm 0.116$                   | 0.830             | <b>1.45</b> M | 12.9 s          |
| Modified U-Net + post-proc. (ours) | <b><math>0.930 \pm 0.013</math></b> | $0.817 \pm 0.017$ | 2.21 M        | 18.7 s          |

patches and (ii) the post-processing block (Algorithm 1) on a CPU. The total inference time per Gavet image (resp. per Alizarine image) is well within the latency budget of an on-demand clinical reporting application, and can be reduced further with a coarser inference stride (see below).

Table 5: Inference time per image on a single NVIDIA A100 40GB (network) and a single CPU core (post-processing). Sliding-window stride 1 px horizontal / 2 px vertical (Sec. 3.5); 10 images  $\times$  5 repeats with 2 warm-up runs; coefficient of variation  $< 1\%$ .

| Stage                                 | Gavet ( $152 \times 388$ ) | Alizarine ( $768 \times 576$ ) |
|---------------------------------------|----------------------------|--------------------------------|
| Network forward pass (sliding window) | $18\,708 \pm 70$ ms        | $195\,414 \pm 430$ ms          |
| Post-processing (CPU)                 | $4.0 \pm 0.2$ ms           | $38.4 \pm 1.8$ ms              |
| <b>Total</b>                          | <b>18.71 s</b>             | <b>195.45 s</b>                |

The forward-pass time is dominated by the dense sliding window (stride  $1 \times 2$  pixels): about 18 000 patches per Gavet image and 191 000 per Alizarine image are individually processed line-by-line for memory reasons. Selecting a larger stride at inference (e.g.  $4 \times 4$ ) reduces the forward-pass time by an order of magnitude (Gavet  $\sim 2.5$  s, Alizarine  $\sim 26$  s) with a marginal Dice drop, providing a tunable trade-off for clinical reporting and batch deployment.

#### 4.2.5 Network performance on Gavet test set

On the Gavet test set, the proposed method achieved a high mean Dice coefficient ( $0.93 \pm 0.01$ ), indicating strong global agreement between predicted and reference cell contours. This performance is consistent with, and in some cases exceeds, previously reported results obtained using U-Net-based architectures on the same dataset (Table 6). The low variability observed across test images further suggests that the segmentation performance is stable despite local variations in contrast and cell density. While the Modified Hausdorff Distance (MHD) exhibited a higher dispersion ( $17.50 \pm 9.39$  pixels), this behavior is expected given its sensitivity to localized boundary errors, particularly at image peripheries. Importantly, these local discrepancies did not substantially affect the overall segmentation quality or the subsequent extraction of morphometric parameters.

| Reference                          | Method                          | Mean DICE       |
|------------------------------------|---------------------------------|-----------------|
| Kucharski & Fabijańska (2023) [21] | U-Net (real data)               | 0.77 - 0.81     |
| Fabijańska (2019) [12]             | Reduced U-Net + post-processing | 0.86            |
| This work                          | U-Net + post-processing         | $0.93 \pm 0.01$ |

Table 6: Comparison of reported segmentation performances on the Gavet dataset.

A key contribution of this work lies in the explicit integration of a post-processing stage tailored to corneal endothelial morphology. While convolutional neural networks provide a powerful basis for boundary detection, the raw segmentation maps produced by the network often contain artifacts that are incompatible with reliable morphometric analysis. The morphological refinement steps applied in this study—adaptive thresholding, component filtering, skeletonization, and pruning—significantly

improved contour continuity and topological consistency. The comparison with previously published methods suggests that this post-processing pipeline plays a central role in achieving high segmentation accuracy, particularly in regions of low contrast or high cell density.

#### 4.2.6 Positioning with respect to recent state-of-the-art methods

To position our framework within the current literature on deep-learning segmentation of the corneal endothelium, Table 7 summarises representative recent methods together with the family of architecture they rely on and the segmentation performance reported by their authors. Beyond the two U-Net baselines evaluated on the same Gavet data (Table 6) and the transformer architectures we retrained ourselves (Swin-UNet and TransUNet, Table 4), recent work has explored lightweight hybrid models (MOBILE-CELLNET [19]), integrated segmentation-and-morphometry pipelines (CELLSDEEPNET [13]), feedback/attention-augmented U-Nets and signed-distance regression for pathological tissue [40, 39], and data-efficient strategies based on synthetic [21] or warping-based [43] augmentation. Two observations emerge from this comparison. First, our compact CNN combined with a topology-aware post-processing pipeline attains a Gavet Dice (0.93) that is competitive with the reported values (which span roughly 0.77–0.98 depending on the dataset, the Dice definition and the ROI convention) while remaining lightweight (2.21 M parameters). Second, and more importantly, most of these methods report *pixel-wise* accuracy only, whereas our framework additionally validates the downstream *morphometric* parameters (ECD, CV, HEX) and their agreement with ground truth under domain shift (Sec. 4.2.8), which is the property that ultimately matters for clinical use. Because the methods of Table 7 were evaluated on different datasets, image modalities and ROI conventions, the reported figures are not directly comparable and should be read as an indication of the operating range rather than as a strict ranking.

Table 7: Positioning with respect to recent deep-learning methods for corneal endothelial segmentation. Reported performance is taken from the original publications and was obtained on different datasets and protocols; values are therefore *not* directly comparable across rows. “–” indicates that the work does not report a Dice coefficient in a comparable form; the metric it does report is, respectively, 83.8% pixel accuracy for [39], a 4.06% mean absolute error on endothelial cell density for MOBILE-CELLNET [19], and sensitivity and specificity both above 0.90 for CELLSDEEPNET [13]. <sup>†</sup>Cell-edge Dice (the same group also reports a cell-body Dice of 0.92).

| Method (year)                                      |           | Family                       | Dataset / modality     | Reported Dice      |
|----------------------------------------------------|-----------|------------------------------|------------------------|--------------------|
| Fabijańska (2018) [12]                             |           | U-Net + post-proc.           | Alizarine              | 0.86               |
| Vigueras-Guillén <i>et al.</i> (2019) [41]         | <i>et</i> | U-Net / SW-net               | specular               | 0.98               |
| Vigueras-Guillén <i>et al.</i> (2022) [40]         | <i>et</i> | DenseUNet (feedback + attn.) | specular + guttae      | 0.86 <sup>†</sup>  |
| Vigueras-Guillén <i>et al.</i> (2023) [39]         | <i>et</i> | Signed-distance regression   | Fuchs (all stages)     | –                  |
| Kucharski & Fabijańska (2023) [21]                 |           | GAN-synthetic + U-Net        | Alizarine/Gavet (real) | 0.77–0.81          |
| Karmakar <i>et al.</i> (2023, MOBILE-CELLNET) [19] |           | Lightweight hybrid CNN       | specular               | –                  |
| Al-Fahdawi <i>et al.</i> (2022, CELLSDEEPNET) [13] |           | CNN + morphometry            | specular               | –                  |
| Wang <i>et al.</i> (2024) [43]                     |           | Semi-sup. + warping aug.     | specular               | 0.94               |
| Swin-UNet [5] (retrained here)                     |           | Transformer                  | Gavet                  | 0.80               |
| TransUNet [7] (retrained here)                     |           | Transformer                  | Gavet                  | 0.82               |
| <b>This work</b>                                   |           | Compact U-Net + post-proc.   | Gavet / Alizarine      | <b>0.93</b> / 0.82 |

#### 4.2.7 Network performance on unseen data of Alizarine dataset

The evaluation on the external Alizarine dataset provides insight into the generalization capabilities of the proposed approach. As expected, segmentation performance decreased relative to the training domain, with a mean Dice coefficient of approximately 0.81 (Figure 9). This reduction reflects the substantial domain shift between datasets, including differences in acquisition conditions, resolution, contrast, and tissue origin. Nevertheless, the obtained performance remains within the upper range of cross-dataset results reported in the literature for models trained exclusively on the Gavet dataset, indicating that the proposed framework retains a coherent representation of endothelial cell structure even under previously unseen imaging conditions. In terms of local boundary accuracy, the method achieved a mean Modified Hausdorff Distance (MHD) of  $9.4 \pm 2.0$  pixels on Alizarine (Figure 10), to be compared with  $17.5 \pm 9.4$  pixels on the smaller Gavet images. Although this metric exhibits higher variability due to its sensitivity to localized contour discrepancies, particularly in regions of dense or low-contrast cells, the observed MHD values remain within an acceptable range for downstream morphometric analysis.

The post-processing pipeline proved particularly important in this cross-dataset setting. While the overall structure of the pipeline was preserved, minor adjustments to the parameters of the adaptive thresholding step were required to account for the lower and more heterogeneous contrast of the Alizarine images. This highlights an important aspect of the proposed approach: although the segmentation network captures generic morphological features, domain-specific image characteristics can still influence the quality of the final contours. Post-processing therefore acts as a compensatory mechanism that enhances robustness when transferring the model to new datasets.

#### Relationship between Dice and Modified Hausdorff Distance

To complement the dataset-level statistics reported above, we examined the per-image relationship between Dice and MHD on both Gavet and Alizarine (Figure 11). As expected, the two metrics are negatively correlated: images with a higher Dice tend to have a lower MHD. However, the correlation is not perfect, which confirms that the two metrics are complementary rather than redundant. In particular, a few Alizarine images with intermediate Dice values ( $\sim 0.78$ – $0.80$ ) exhibited substantially larger MHD values, reflecting the presence of localized boundary errors (typically near the periphery of the focused region) that have a limited impact on global overlap but a disproportionate effect on the worst-case localization error. The detailed limitations of the proposed approach—including the small dataset size, the empirical tuning of post-processing parameters, the evaluation on healthy or near-healthy endothelium, and the sensitivity to image quality—are discussed in Section 5 just before the conclusion.

Such reliable edge detection forms a crucial foundation for the subsequent step: the extraction of morphometric parameters. Accurate contours are essential for determining cell density, hexagonality, and cell size variation, ensuring that morphometric measurements remain robust and comparable across different datasets.

#### 4.2.8 Determination of morphometric parameters

Cellular morphometric parameters of the corneal endothelium were automatically derived from the post-processed segmentation maps generated by the proposed model. Following binary inversion, individual endothelial cells were identified using connected component labeling. Border-touching cells were excluded from the analysis to avoid partial cell bias.

Three key morphometric parameters were computed: endothelial cell density (ECD), coefficient of variation of cell area (CV), and pleomorphism.

1. Endothelial cell density (ECD) represents the number of endothelial cells per square millimeter ( $\text{cells}/\text{mm}^2$ ). To calculate ECD, only fully visible cells within the region of interest were counted, excluding cells touching the image borders to avoid partial cell bias. The total number of valid cells was then divided by the physical area of the analyzed region, taking into account the calibrated pixel dimensions. ECD reflects the functional status of the corneal endothelium, with decreased density often associated with endothelial dysfunction or disease.

2. The coefficient of variation (CV) quantifies the variability of individual cell sizes. It is defined as:

$$CV = \frac{\text{standard deviation of cell areas}}{\text{mean cell area}} \times 100$$

A low CV value indicates a homogeneous cell population, while a high CV value reflects polymorphism, a condition in which cell sizes vary considerably. CV assessment is particularly sensitive to segmentation accuracy, as errors in cell boundary delineation can artificially increase variability.

3. Pleomorphism measures the percentage of morphologically regular cells. A cell was considered regular if its shape approximated an ideal hexagon. To quantify this, a regularity score was calculated for each cell based on three simple shape descriptors:

**Circularity** Circularity measures the compactness of the cell.

$$\text{Circularity} = \frac{4\pi \times \text{cell area}}{\text{perimeter}^2}$$

A value of 1 corresponds to a perfect circle; lower values indicate irregular shapes.

**Solidity** Solidity evaluates contour regularity.

$$\text{Solidity} = \frac{\text{cell area}}{\text{area of convex hull}}$$

A solidity close to 1 indicates a smooth, well-defined contour, while lower values reflect concavities or irregular borders.

**Aspect ratio** Aspect ratio captures shape symmetry.

$$\text{Aspect ratio} = \frac{\text{major axis length}}{\text{minor axis length}}$$

Values near 1 indicate isotropic cells, whereas higher values correspond to elongated shapes.

These three descriptors were combined into a single regularity score:

$$\text{Regularity score} = \frac{\text{Circularity} + \text{Solidity} + \frac{1}{\text{Aspect ratio}}}{3}$$

Cells with a regularity score above a predefined threshold were classified as morphologically regular. Pleomorphism was then expressed as the percentage of regular cells among all analyzed cells. This parameter provides a quantitative measure of endothelial cell shape uniformity, which is important for assessing corneal health.

To assess the agreement between the predicted morphometric parameters and the ground truth values derived from the reference masks, analyses were conducted on two datasets: the Gavet test set, which originates from the training domain, and the Alizarine dataset, which was not seen during training. For each image, errors were computed as the difference between the post-processed predictions and the corresponding ground truth values for endothelial cell density, coefficient of variation (CV), and pleomorphism, along with the relative percentage error for cell density.

Because the measurements consist of paired observations (predictions and reference values for the same image) and the distributions of the differences are not guaranteed to be normal—particularly for CV and pleomorphism, which are sensitive to segmentation variability—the Wilcoxon signed-rank test was employed instead of a paired t-test. This non-parametric test does not assume normality and is therefore better suited for small sample sizes and datasets containing outliers. It evaluates whether the median difference between predicted and ground truth values significantly deviates from zero, allowing the identification of systematic biases.

The statistical analysis was performed separately for each morphometric parameter—cell count, cell density, coefficient of variation, and pleomorphism—to evaluate both the significance of observed

differences and the overall agreement between values computed from the post-processed predictions and those obtained directly from the ground truth masks. Model performance on both the Gavet and Alizarine datasets was summarized using the Wilcoxon signed-rank test (Table 8) and complementary error metrics (Table 9).

| Parameter    | Dataset   | p-value               |
|--------------|-----------|-----------------------|
| Cells        | Gavet     | $3.62 \times 10^{-5}$ |
| Cells        | Alizarine | $2.49 \times 10^{-6}$ |
| Density      | Gavet     | $3.26 \times 10^{-5}$ |
| Density      | Alizarine | $3.73 \times 10^{-9}$ |
| CV           | Gavet     | $7.46 \times 10^{-3}$ |
| CV           | Alizarine | $3.73 \times 10^{-9}$ |
| Pleomorphism | Gavet     | $3.73 \times 10^{-9}$ |
| Pleomorphism | Alizarine | $3.73 \times 10^{-9}$ |

Table 8: Wilcoxon signed-rank test for predicted vs ground-truth values.

The Wilcoxon signed-rank test revealed statistically significant differences between model predictions and ground-truth measurements for all parameters across both datasets. However, the magnitude and direction of errors differ between the training/test data and the unseen dataset.

| Parameter    | Dataset   | Mean  | SD   | Min   | Max   |
|--------------|-----------|-------|------|-------|-------|
| Density (%)  | Gavet     | -2.09 | 2.01 | -5.88 | 2.86  |
| CV           | Gavet     | 0.87  | 2.47 | -4.79 | 8.30  |
| Pleomorphism | Gavet     | 7.46  | 2.70 | 2.42  | 13.31 |
| Density (%)  | Alizarine | -2.94 | 1.67 | -7.13 | -0.71 |
| CV           | Alizarine | -0.18 | 0.11 | -0.48 | -0.02 |
| Pleomorphism | Alizarine | 8.30  | 4.74 | 0.22  | 21.39 |

Table 9: Summary of prediction errors.

### Bland–Altman analysis

On the Gavet test set, endothelial cell density and coefficient of variation exhibited low bias and narrow limits of agreement with ground-truth measurements, as demonstrated by Bland–Altman analysis (Figure 12). These results indicate that the proposed framework enables accurate and reproducible estimation of clinically relevant parameters when applied to data drawn from the same distribution as the training set. Pleomorphism, while slightly overestimated, remained within acceptable agreement limits, reflecting the inherent sensitivity of shape-based metrics to minor segmentation errors.

On the Alizarine dataset, larger biases and wider limits of agreement were observed, particularly for pleomorphism. While cell density tended to be systematically underestimated, the magnitude of this bias remained moderate. In contrast, pleomorphism showed increased variability and overestimation, underscoring its strong dependence on precise boundary delineation and cell shape regularity. These findings are consistent with previous reports indicating that shape-related parameters are more sensitive to segmentation imperfections than size- or count-based metrics. Importantly, despite statistically significant differences detected by the Wilcoxon tests, Bland–Altman analysis (Figure 12) demonstrated that most errors remained within ranges considered acceptable for practical or clinical interpretation.

**Clinical interpretation of the Bland–Altman limits of agreement.** For ECD, the bias is approximately  $-2\%$  on Gavet and  $-3\%$  on Alizarine, i.e. in absolute terms a few tens of cells/mm<sup>2</sup>. Since clinically meaningful changes in ECD are usually of the order of several hundreds of cells/mm<sup>2</sup> between healthy ( $\sim 2,200$ – $3,000$  cells/mm<sup>2</sup>) and pathological ( $< 2,000$  cells/mm<sup>2</sup>) endothelium [3, 24], the systematic underestimation observed here remains well below clinical decision thresholds, and the limits of agreement do not extend across these thresholds. For CV, the mean bias is low ( $< 1\%$  on both datasets), indicating excellent overall agreement; this average error is minor compared to the 10% gap separating healthy thresholds (30%) from severe polymegathism (40%) (see Sec. 3.7). However, the maximum overestimation reached 8.30% on the Gavet dataset, suggesting that isolated segmentation

inaccuracies can noticeably affect individual CV estimates for borderline cases near the 30% clinical threshold. For pleomorphism (HEX), the bias is larger ( $\sim 7\text{--}8\%$ ) and the limits of agreement are wider, which is consistent with HEX being a shape-based parameter highly sensitive to local segmentation errors. The clinical implication is that the proposed framework can be used directly for automated estimation of ECD and CV, while for HEX the absolute value should be interpreted with caution and is more reliably used to monitor relative changes within a patient than to make absolute comparisons across patients or devices.

Despite these limitations, the proposed framework demonstrates that combining a tailored deep learning architecture with domain-specific post-processing enables accurate and reproducible segmentation of corneal endothelial cells, as well as reliable estimation of key morphometric parameters. These results support the potential of the approach as a practical tool for automated endothelial assessment and provide a solid foundation for future extensions toward more heterogeneous datasets and pathological conditions.

## 5 Limitations

Although the proposed framework achieves accurate and reproducible segmentation as well as reliable morphometric quantification on both the Gavet test set and the external Alizarine dataset, several limitations must be acknowledged in order to delineate its current scope of applicability and to motivate future work.

**Training-data size and diversity.** The model was trained on a single dataset of 30 specular-microscopy images of healthy human corneal endothelium (Gavet). Although data augmentation (horizontal/vertical symmetries) and a high patch density partially mitigate this limitation, 30 images cannot fully capture the variability of acquisition devices, illumination conditions, magnifications, ethnicities or pathologies encountered in clinical practice. As a consequence, our framework is best regarded as a robust prototype rather than as a clinically validated tool.

**Generalization to images of different modality.** Our cross-dataset evaluation focused on Alizarine, which differs from Gavet in resolution, tissue (porcine), and stain (alizarine red). The framework remained robust under this domain shift, but generalization to other modalities (e.g. confocal microscopy, optical coherence tomography), to wide-field instruments, or to other staining protocols was not explicitly evaluated and should be the subject of further studies.

**Sensitivity to image quality and out-of-focus regions.** Specular-microscopy images frequently contain out-of-focus areas, motion blur, partial occlusions or stain artefacts—issues regularly encountered by clinicians in routine practice. The proposed framework relies on the underlying network being able to detect cell boundaries from the input image; in regions where the boundary evidence is degraded by noise, blur or stain residues, the raw probability map becomes unreliable and the post-processing pipeline cannot recreate cells whose evidence is missing from the input. As a consequence, the most robust morphometric estimates are obtained when the analysis is restricted to a well-focused region of interest, as we did in our evaluation. Performance on systematically low-quality or out-of-focus images is expected to be worse than the values reported here, and a quality-control upstream module—automatically rejecting unusable images or restricting the analysis to in-focus sub-regions—would be a useful addition before clinical deployment.

**Empirical post-processing parameters.** Three parameters of the post-processing pipeline (Sauvola window size  $w$ , sensitivity  $k$ , pruning length  $\ell$ ) had to be re-tuned manually when transferring from Gavet to Alizarine. Although the qualitative effect of each parameter is well understood, this manual tuning hinders out-of-the-box deployment on a new device. Automating this step—through grid search on a small calibration set, Bayesian optimization, or learning the post-processing parameters jointly with the network—is a clear direction for future work.

**Hyperparameter and architectural sensitivity.** Although our ablation study (Sec. 4.2.1) and the comparison with transformer-based architectures (Sec. 4.2.3) suggest that the precise architecture

is not critical when combined with our post-processing, a more thorough sensitivity analysis (patch size, stride, number of filters, dropout rate, choice of loss) would further support these design decisions. Hyperparameter optimization was performed manually rather than through systematic search.

**Pathological cases.** The Gavet dataset consists primarily of healthy or near-healthy corneal endothelium. Pathological cases such as advanced Fuchs’ dystrophy with extensive guttae, corneal grafts after DMEK/DSAEK, or post-traumatic endothelium were not represented in the training data. Pleomorphism, the parameter most sensitive to local segmentation errors in our experiments, is also the parameter most affected by these pathologies; further work will therefore focus on training and evaluating the framework on pathological datasets such as those used in [40, 6, 10].

**Computational footprint and scalability.** Although the network itself is lightweight ( $\sim 2.2$  M parameters) and the sliding-window inference is memory-efficient by design, processing very large widefield images would still require additional engineering (tile-based inference, GPU/CPU pipelining). Scaling the framework to widefield clinical instruments and to clinical batches of hundreds of images per session is therefore a practical challenge that would benefit from further system-level optimization.

**Clinical validation.** Finally, we did not perform a prospective clinical validation comparing automated outputs to clinical decisions made by ophthalmologists. Such a study, ideally including inter-observer variability among experts as a baseline, is the next step to translate the proposed framework into a clinically actionable tool.

## 6 Conclusion

This study presented a fully automated framework for corneal endothelial cell segmentation and morphometric analysis, combining a lightweight U-Net architecture with a dedicated morphological post-processing pipeline. By explicitly integrating domain-specific refinement steps into the segmentation workflow, the proposed approach produces topologically consistent cell contours suitable for reliable morphometric quantification.

Experimental results demonstrated high segmentation accuracy on the Gavet dataset, with a mean Dice coefficient of 0.93, and stable estimation of clinically relevant parameters such as endothelial cell density and coefficient of variation. The evaluation on an external unseen dataset further confirmed the robustness of the method under domain shift, with performance remaining comparable to the upper range of cross-dataset results reported in the literature. While pleomorphism estimation proved more sensitive to segmentation imperfections, the observed discrepancies remained within acceptable limits for practical interpretation.

## Theoretical and practical implications

From a methodological standpoint, our results suggest that, when the training set is small and the targeted clinical outputs are derived from cell topology rather than from pixel-wise labels, a lightweight CNN combined with a topology-aware post-processing pipeline can match or surpass transformer-based architectures of comparable or larger size (Swin-UNet, TransUNet). This challenges the implicit assumption that performance gains in medical image segmentation always come from larger and more complex models: in our setting, the bottleneck is the topological consistency of the predicted contours rather than the raw pixel-wise accuracy of the network. Practically, our framework is efficient (a few seconds per image on a single GPU, reducible by roughly an order of magnitude with a coarser inference stride, see Table 5), reproducible, and produces ECD and CV measurements with biases below the inter-observer variability reported in the clinical literature, which makes it a credible candidate for routine deployment in research and clinical settings.

## Research contributions

This work makes the following contributions to the field of automated corneal endothelial assessment:

- A *lightweight U-Net* explicitly designed for small specular-microscopy patches and small training sets, with a clear and reproducible specification of the architectural modifications relative to the vanilla U-Net.
- A *topology-aware post-processing pipeline* (Algorithm 1) whose every step is documented, parameterized and reproducible, thereby filling a recurrent gap in the literature where post-processing is often only briefly sketched.
- An *end-to-end morphometric pipeline* (Algorithm 2) that derives ECD, CV and HEX directly from the post-processed contours, including a quantitative regularity score.
- A *cross-dataset evaluation* on Gavet and Alizarine that quantifies, for the first time on this benchmark, the joint impact of domain shift on segmentation metrics (Dice, MHD), clinical metrics (ECD, CV, HEX) and per-image agreement (Bland–Altman).
- An *ablation study* and a *comparison with state-of-the-art transformer-based architectures* which jointly demonstrate that the proposed combination of lightweight CNN and topology-aware post-processing is competitive with or superior to transformer-based alternatives in this regime.

## Practical advantages

The proposed framework offers several advantages from a practical and clinical perspective. It is *lightweight*: the network has approximately 2.2 million parameters (about  $3\times$  fewer than Swin-UNet), is trained in  $\sim 10$  minutes on a single A100 GPU and runs inference in seconds to a couple of minutes per image depending on image size and sliding-window stride (Table 5), making it compatible with clinic-side batch reporting. It is *interpretable*: every step of the post-processing pipeline—thresholding, morphological cleanup, skeletonization, pruning—has a clear mathematical definition and a visualizable intermediate output, which is critical for clinical acceptance and for failure analysis. It is *reproducible*: the data, preprocessing, training, inference and post-processing parameters are explicitly documented and the only dataset-specific tuning concerns three intuitive parameters of the thresholding/skeletonization steps. It is *robust to domain shift*: cross-dataset Dice on Alizarine remains in the upper range of values reported in the literature for models trained exclusively on Gavet, and the resulting biases on ECD and CV are well below clinical decision thresholds. Finally, it is *morphometry-aware*: the design priority is not raw pixel-wise accuracy but the reliability of the clinical parameters derived from the segmentation, which is the actual quantity of interest in routine practice.

## Research limitations

A number of limitations, summarized in Section 5, must be kept in mind when interpreting our results. The training set is small (30 images) and limited to healthy or near-healthy human corneal endothelium acquired with one specific device; cross-dataset evaluation was performed on a single external dataset (Alizarine), which itself differs from Gavet in tissue, staining and acquisition conditions but does not cover the full diversity of clinical instruments. Three parameters of the post-processing pipeline had to be re-tuned manually when transferring to Alizarine, and the framework currently lacks an automated upstream image-quality control module: low-quality or out-of-focus images are not automatically rejected, which is an issue regularly encountered in clinical routine. Finally, no prospective clinical validation comparing the automated outputs to expert clinical decisions has yet been performed, and therefore conclusions on clinical actionability remain preliminary.

## Future research directions

Building on these contributions and limitations, we identify three concrete directions for future research:

1. **Validation on pathological cases and multi-device datasets.** Extending training and evaluation to pathological corneas (advanced Fuchs’ dystrophy with guttae, post-DMEK/DSAEK grafts, post-traumatic endothelium) and to widefield specular microscopy [6, 10], in collaboration with clinical partners, would clarify the true clinical applicability of the framework and would address the most pressing limitation of the present study.

2. **Near-real-time integration in the clinical workflow with human-in-the-loop correction.** Integrating the framework directly into specular-microscope acquisition software would enable near-real-time automated reporting of ECD, CV and HEX shortly after acquisition. Coupling this with a Human-in-the-Loop (HITL) correction interface—in which the clinician can validate or locally edit the automatically produced contours—would both enhance clinical trust and continuously enrich the training set, leading to a self-improving system over time.
3. **Adaptive and learnable post-processing.** Replacing the manual tuning of the post-processing parameters by a data-driven mechanism (e.g. Bayesian optimization on a small calibration set, or differentiable post-processing trained jointly with the network) would remove the only remaining piece of dataset-specific configuration and pave the way for plug-and-play deployment on new devices and modalities. Combining this with an automated image-quality assessment module would directly tackle the low-quality / out-of-focus issue identified in Section 5.

Overall, the proposed framework provides a solid and reproducible basis for automated corneal endothelial assessment and represents a step toward reliable integration of artificial intelligence-based tools into routine ophthalmic practice.

## Acknowledgements

This work was granted access to the HPC resources of IDRIS under the allocations [AD010613582R4] made by GENCI. We also thank the Mesocentre of Franche-Comté for the computing resources. This work was also supported by the EIPHI Graduate School (contract No. ANR-17-EURE-0002).

## References

- [1] J. Angulo and S. Matou. “Automatic quantification of in vitro endothelial cell networks using mathematical morphology”. In: *Proceedings of the 5th IASTED International Conference on Visualization, Imaging, and Image Processing (VIIP’05)*. 2005, pp. 51–56.
- [2] G. Ayala, M.E. Díaz, and L. Martínez-Costa. “Granulometric moments and corneal endothelium status”. In: *Pattern Recognition* 34.6 (2001), pp. 1219–1227. ISSN: 0031-3203. DOI: [https://doi.org/10.1016/S0031-3203\(00\)00074-1](https://doi.org/10.1016/S0031-3203(00)00074-1). URL: <https://www.sciencedirect.com/science/article/pii/S0031320300000741>.
- [3] W. Bourne. “Biology of the corneal endothelium in health and disease”. In: *Eye* 17 (2003), pp. 912–918. DOI: [10.1038/sj.eye.6700559](https://doi.org/10.1038/sj.eye.6700559).
- [4] Julien Bullet et al. “A reproducible automated segmentation algorithm for corneal epithelium cell images from in vivo laser scanning confocal microscopy”. In: *Acta Ophthalmologica* 92.4 (2014), e312–e316. DOI: <https://doi.org/10.1111/aos.12304>. eprint: <https://onlinelibrary.wiley.com/doi/pdf/10.1111/aos.12304>. URL: <https://onlinelibrary.wiley.com/doi/abs/10.1111/aos.12304>.
- [5] H. Cao et al. “Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation”. In: *arXiv preprint arXiv:2105.05537* (2021).
- [6] F. Cassinero et al. “Deep learning-assisted widefield endothelial imaging in Descemet membrane endothelial keratoplasty”. In: *Frontiers in Medicine* 12 (2025). DOI: [10.3389/fmed.2025.1715673](https://doi.org/10.3389/fmed.2025.1715673). URL: <https://www.frontiersin.org/journals/medicine/articles/10.3389/fmed.2025.1715673/full>.
- [7] J. Chen et al. “TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation”. In: *arXiv preprint arXiv:2102.04306* (2021).
- [8] D. Cireşan et al. “Deep neural networks segment neuronal membranes in electron microscopy images”. In: *Advances in Neural Information Processing Systems* 25. 2012, pp. 2843–2851.
- [9] Issam Dagher and Kamal El Tom. “WaterBalloons: A hybrid watershed Balloon Snake segmentation”. In: *Image and Vision Computing* 26.7 (2008), pp. 905–912. ISSN: 0262-8856. DOI: <https://doi.org/10.1016/j.imavis.2007.10.010>. URL: <https://www.sciencedirect.com/science/article/pii/S0262885607001965>.

- [10] K. Daniels-Tomaczak et al. “Deep Learning Analysis of Widefield Cornea Endothelial Imaging in Fuchs Dystrophy”. In: *Ophthalmology Science* (2025). TODO: complete authors and verify DOI/pages. DOI: [10.1016/j.xops.2025.100212](https://doi.org/10.1016/j.xops.2025.100212). URL: [https://www.ophtalmologyscience.org/article/S2666-9145\(25\)00212-X/fulltext](https://www.ophtalmologyscience.org/article/S2666-9145(25)00212-X/fulltext).
- [11] Anna Fabijańska. “Corneal Endothelium Image Segmentation Using Feedforward Neural Network”. In: Sept. 2017, pp. 629–637. DOI: [10.15439/2017F54](https://doi.org/10.15439/2017F54).
- [12] Anna Fabijańska. “Segmentation of corneal endothelium images using a U-Net-based convolutional neural network”. In: *Artificial Intelligence in Medicine* 88 (2018), pp. 1–13. ISSN: 0933-3657. DOI: <https://doi.org/10.1016/j.artmed.2018.04.004>. URL: <https://www.sciencedirect.com/science/article/pii/S0933365718300575>.
- [13] S. Al-Fahdawi et al. “CellsDeepNet: A Novel Deep Learning-Based Web Application for the Automated Morphometric Analysis of Corneal Endothelial Cells”. In: *Mathematics* 10.3 (2022), p. 320. DOI: [10.3390/math10030320](https://doi.org/10.3390/math10030320).
- [14] Marco Foracchia and Alfredo Ruggeri. “Corneal Endothelium Cell Field Analysis by means of Interacting Bayesian Shape Models”. In: *2007 29th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. 2007, pp. 6035–6038. DOI: [10.1109/IEMBS.2007.4353724](https://doi.org/10.1109/IEMBS.2007.4353724).
- [15] Y. F. Garcia-Marin et al. “Recent advances in corneal specular microscopy image analysis through artificial intelligence”. In: *PLOS Digital Health* (2026). TODO: complete authors and verify metadata. DOI: [10.1371/journal.pdig.0001305](https://doi.org/10.1371/journal.pdig.0001305).
- [16] Krzysztof Habrat et al. “Cell Detection in Corneal Endothelial Images Using Directional Filters”. In: *Image Processing and Communications Challenges 7*. Ed. by Ryszard S. Choraś. Cham: Springer International Publishing, 2016, pp. 113–123. ISBN: 978-3-319-23814-2.
- [17] Raidel Herrera-Pereda et al. “A Review on Digital Image Processing Techniques for In-Vivo Confocal Images of the Cornea”. In: *Medical Image Analysis* 73 (July 2021), p. 102188. DOI: [10.1016/j.media.2021.102188](https://doi.org/10.1016/j.media.2021.102188).
- [18] N. Joseph et al. “Current Applications of Artificial Intelligence for Fuchs Endothelial Corneal Dystrophy: A Systematic Review”. In: *Translational Vision Science & Technology* (2025). TODO: complete authors and verify metadata.
- [19] R. Karmakar, S. Nooshabadi, and A. O. Eghrari. “Mobile-CellNet: Automatic Segmentation of Corneal Endothelium Using an Efficient Hybrid Deep Learning Model”. In: *Cornea* 42.4 (2023), pp. 456–463. DOI: [10.1097/IC0.00000000000003186](https://doi.org/10.1097/IC0.00000000000003186).
- [20] Saya Katafuchi and Motohide Yoshimura. “Convolution neural network for contour extraction of corneal endothelial cells”. In: *Thirteenth International Conference on Quality Control by Artificial Vision 2017*. Ed. by Hajime Nagahara, Kazunori Umeda, and Atsushi Yamashita. Vol. 10338. International Society for Optics and Photonics. SPIE, 2017, p. 103380L. DOI: [10.1117/12.2264430](https://doi.org/10.1117/12.2264430). URL: <https://doi.org/10.1117/12.2264430>.
- [21] Adrian Kucharski and Anna Fabijańska. “Corneal endothelial image segmentation training data generation using GANs. Do experts need to annotate?” In: *Biomedical Signal Processing and Control* 85 (2023), p. 104985. ISSN: 1746-8094. DOI: <https://doi.org/10.1016/j.bspc.2023.104985>. URL: <https://www.sciencedirect.com/science/article/pii/S1746809423004184>.
- [22] T.-Y. Lin et al. “Focal Loss for Dense Object Detection”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42.2 (2020), pp. 318–327. DOI: [10.1109/TPAMI.2018.2858826](https://doi.org/10.1109/TPAMI.2018.2858826).
- [23] M. Mahzoun et al. “Detection and complement of hexagonal borders in corneal endothelial cell image”. In: *Medical Imaging Technology* 14.1 (1996), pp. 56–69.
- [24] B. E. McCarey, H. F. Edelhauser, and M. J. Lynn. “Review of corneal endothelial specular microscopy for FDA clinical trials of refractive procedures, surgical devices, and new intraocular drugs and solutions”. In: *Cornea* 27.1 (2008), pp. 1–16. DOI: [10.1097/IC0.0b013e31815892da](https://doi.org/10.1097/IC0.0b013e31815892da).
- [25] F. Milletari, N. Navab, and S.-A. Ahmadi. “V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation”. In: (2016), pp. 565–571. DOI: [10.1109/3DV.2016.79](https://doi.org/10.1109/3DV.2016.79).

- [26] Ryoichi Nadachi and Kazuo Nunokawa. “Automated corneal endothelial cell analysis”. In: *Fifth Annual IEEE Symposium on Computer-Based Medical Systems (CBMS’92)*, June 14–17, 1992, Durham, NC, USA. IEEE Computer Society, 1992, pp. 450–457. DOI: [10.1109/CBMS.1992.245000](https://doi.org/10.1109/CBMS.1992.245000). URL: <https://doi.org/10.1109/CBMS.1992.245000>.
- [27] Karolina Nurzynska. “Deep Learning as a Tool for Automatic Segmentation of Corneal Endothelium Images”. In: *Symmetry* 10.3 (2018). ISSN: 2073-8994. URL: <https://www.mdpi.com/2073-8994/10/3/60>.
- [28] Karolina Nurzynska. “Problems with Deep Learning Application to Medical Data: Automatic Segmentation of Corneal Endothelium Layer”. In: *Procedia Computer Science* 225 (2023). 27th International Conference on Knowledge Based and Intelligent Information and Engineering Systems (KES 2023), pp. 134–143. ISSN: 1877-0509. DOI: <https://doi.org/10.1016/j.procs.2023.09.100>. URL: <https://www.sciencedirect.com/science/article/pii/S1877050923011559>.
- [29] originlake. *DSE-skeleton-pruning*. <https://github.com/originlake/DSE-skeleton-pruning>. Accessed: 2025-04-29. 2023.
- [30] Enea Poletti and Alfredo Ruggeri. “Segmentation of Corneal Endothelial Cells Contour through Classification of Individual Component Signatures”. In: *IFMBE Proceedings* 41 (Jan. 2014), pp. 411–414. DOI: [10.1007/978-3-319-00846-2\\_102](https://doi.org/10.1007/978-3-319-00846-2_102).
- [31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. *U-Net: Convolutional Networks for Biomedical Image Segmentation*. 2015. arXiv: [1505.04597 \[cs.CV\]](https://arxiv.org/abs/1505.04597). URL: <https://arxiv.org/abs/1505.04597>.
- [32] Alfredo Ruggeri et al. “A system for the automatic estimation of morphometric parameters of corneal endothelium in alizarine red-stained images”. In: *The British journal of ophthalmology* 94 (May 2010), pp. 643–7. DOI: [10.1136/bjo.2009.166561](https://doi.org/10.1136/bjo.2009.166561).
- [33] Francisco J. Sanchez-Marin. “Automatic segmentation of contours of corneal cells”. In: *Computers in Biology and Medicine* 29.4 (1999), pp. 243–258. ISSN: 0010-4825. DOI: [https://doi.org/10.1016/S0010-4825\(99\)00010-4](https://doi.org/10.1016/S0010-4825(99)00010-4). URL: <https://www.sciencedirect.com/science/article/pii/S0010482599000104>.
- [34] J. Sauvola and M. Pietikäinen. “Adaptive document image binarization”. In: *Pattern Recognition* 33.2 (2000), pp. 225–236. ISSN: 0031-3203. DOI: [https://doi.org/10.1016/S0031-3203\(99\)00055-2](https://doi.org/10.1016/S0031-3203(99)00055-2). URL: <https://www.sciencedirect.com/science/article/pii/S0031320399000552>.
- [35] Fabio Scarpa and Alfredo Ruggeri. “Development of a Reliable Automated Algorithm for the Morphometric Analysis of Human Corneal Endothelium”. In: *Cornea* 35.9 (2016), pp. 1222–1228. DOI: [10.1097/ICD.0000000000000908](https://doi.org/10.1097/ICD.0000000000000908).
- [36] B. Selig et al. “Fully automatic evaluation of the corneal endothelium from in vivo confocal microscopy”. In: *BMC Medical Imaging* 15.1 (2015), p. 13. ISSN: 1471-2342. DOI: [10.1186/s12880-015-0054-3](https://doi.org/10.1186/s12880-015-0054-3). URL: <https://doi.org/10.1186/s12880-015-0054-3>.
- [37] T. Romen Singh et al. *A New Local Adaptive Thresholding Technique in Binarization*. 2012. arXiv: [1201.5227 \[cs.CV\]](https://arxiv.org/abs/1201.5227). URL: <https://arxiv.org/abs/1201.5227>.
- [38] B. Van den Bogerd et al. “Corneal Endothelial Cells Over the Past Decade: Are We Missing the Mark(er)?” In: *Translational Vision Science & Technology* 8.6 (Nov. 2019), p. 13. DOI: [10.1167/tvst.8.6.13](https://doi.org/10.1167/tvst.8.6.13). URL: <https://doi.org/10.1167/tvst.8.6.13>.
- [39] J. P. Vigueras-Guillén et al. “Corneal endothelium assessment in specular microscopy images with Fuchs’ dystrophy via deep regression of signed distance maps”. In: *Biomedical Optics Express* (2023). DOI: [10.1364/B0E.477495](https://doi.org/10.1364/B0E.477495).
- [40] J. P. Vigueras-Guillén et al. “DenseUNets with feedback non-local attention for the segmentation of specular microscopy images of the corneal endothelium with guttae”. In: *Scientific Reports* 12.1 (2022), p. 14035. DOI: [10.1038/s41598-022-18180-1](https://doi.org/10.1038/s41598-022-18180-1).
- [41] JP Vigueras-Guillén et al. “Fully convolutional architecture vs sliding-window CNN for corneal endothelium cell segmentation”. In: *BMC Biomed Eng* 1 (2019), p. 4. DOI: [10.1186/s42490-019-0003-2](https://doi.org/10.1186/s42490-019-0003-2). URL: <https://doi.org/10.1186/s42490-019-0003-2>.

- [42] Luc M. Vincent and Barry R. Masters. “Morphological image processing and network analysis of cornea endothelial cell images”. In: *Proc. SPIE 1769, Image Algebra and Morphological Image Processing III*. 1992. DOI: [10.1117/12.60644](https://doi.org/10.1117/12.60644). URL: <https://doi.org/10.1117/12.60644>.
- [43] S. Wang et al. “Data augmentation via warping transforms for modeling natural variability in the corneal endothelium enhances semi-supervised segmentation”. In: *PLOS ONE* 19.11 (2024), e0311849. DOI: [10.1371/journal.pone.0311849](https://doi.org/10.1371/journal.pone.0311849).
- [44] Yinglin Zhang et al. *A Multi-Branch Hybrid Transformer Network for Corneal Endothelial Cell Segmentation*. 2021. arXiv: [2106.07557](https://arxiv.org/abs/2106.07557) [cs.CV]. URL: <https://arxiv.org/abs/2106.07557>.

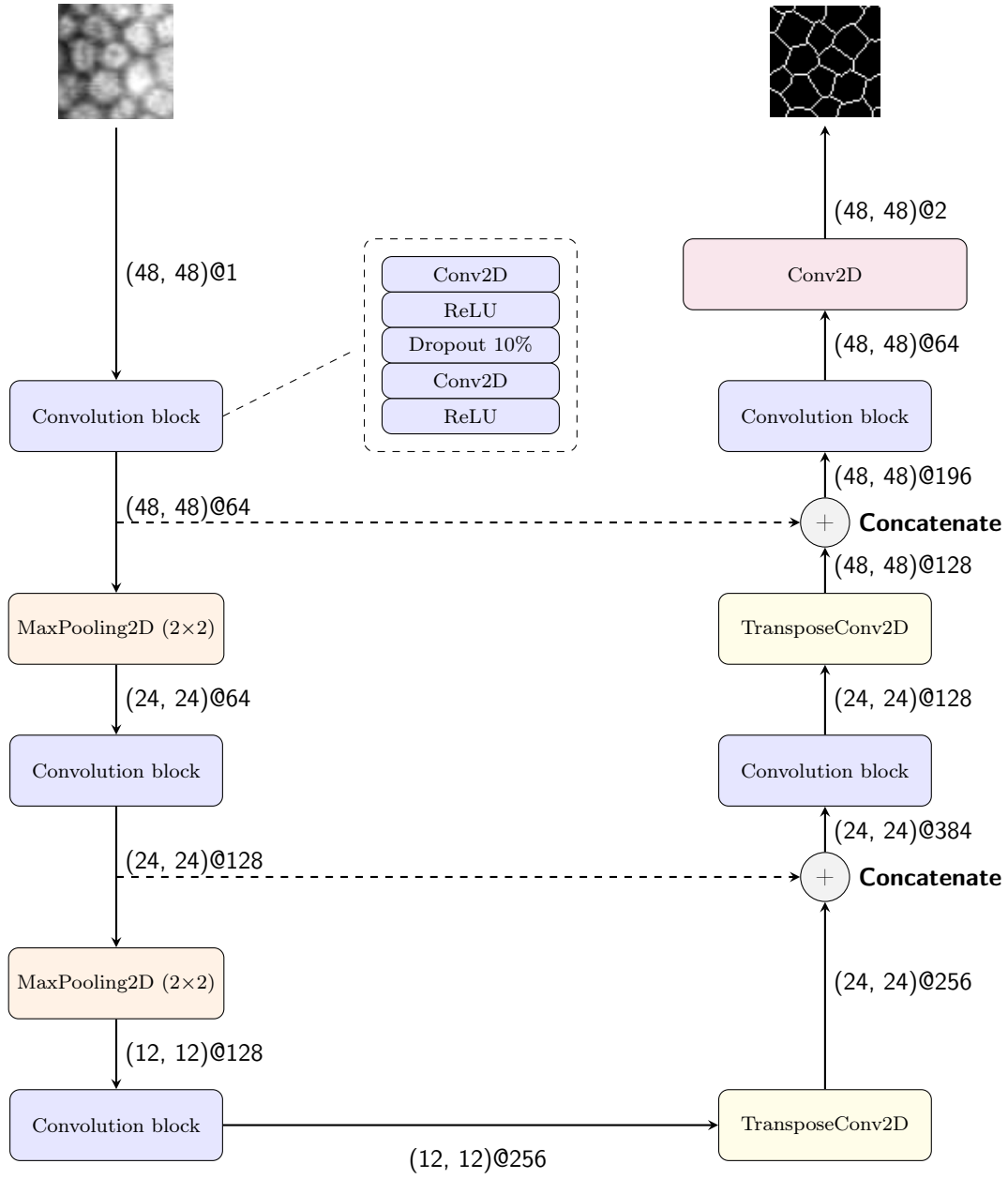


Figure 3: Architecture of the modified U-Net network used in this study. The model takes as input  $48 \times 48$  single-channel (grayscale) image patches. It integrates skip connections via concatenation and downsamples to a  $12 \times 12$  bottleneck with 256 channels. The internal composition of a convolution block is detailed in the center.

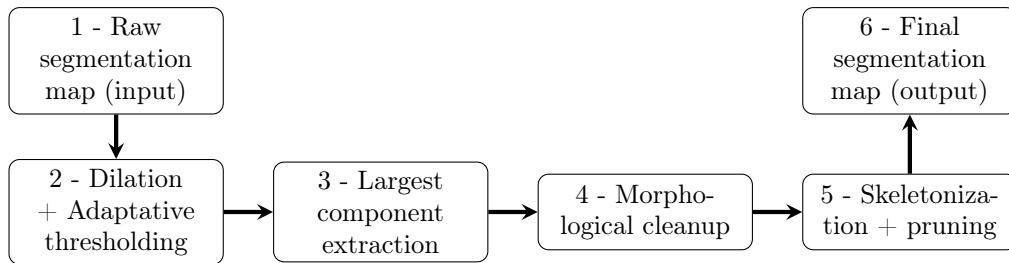


Figure 4: Postprocessing pipeline for image segmentation enhancement

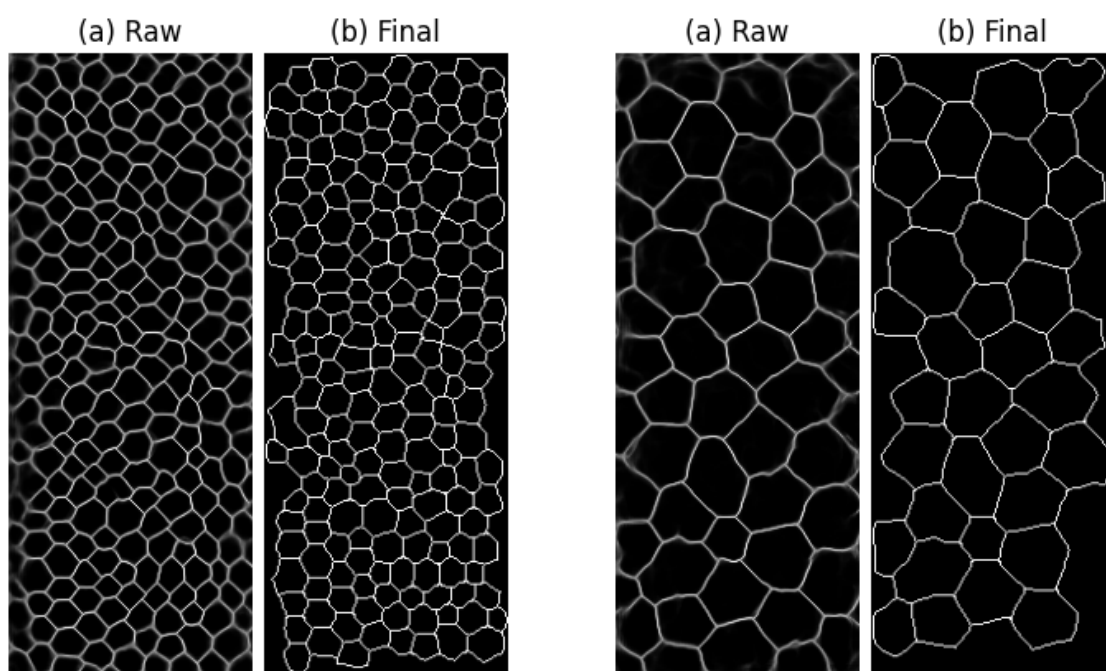


Figure 5: Two examples of post-processing applied to raw segmentation maps.

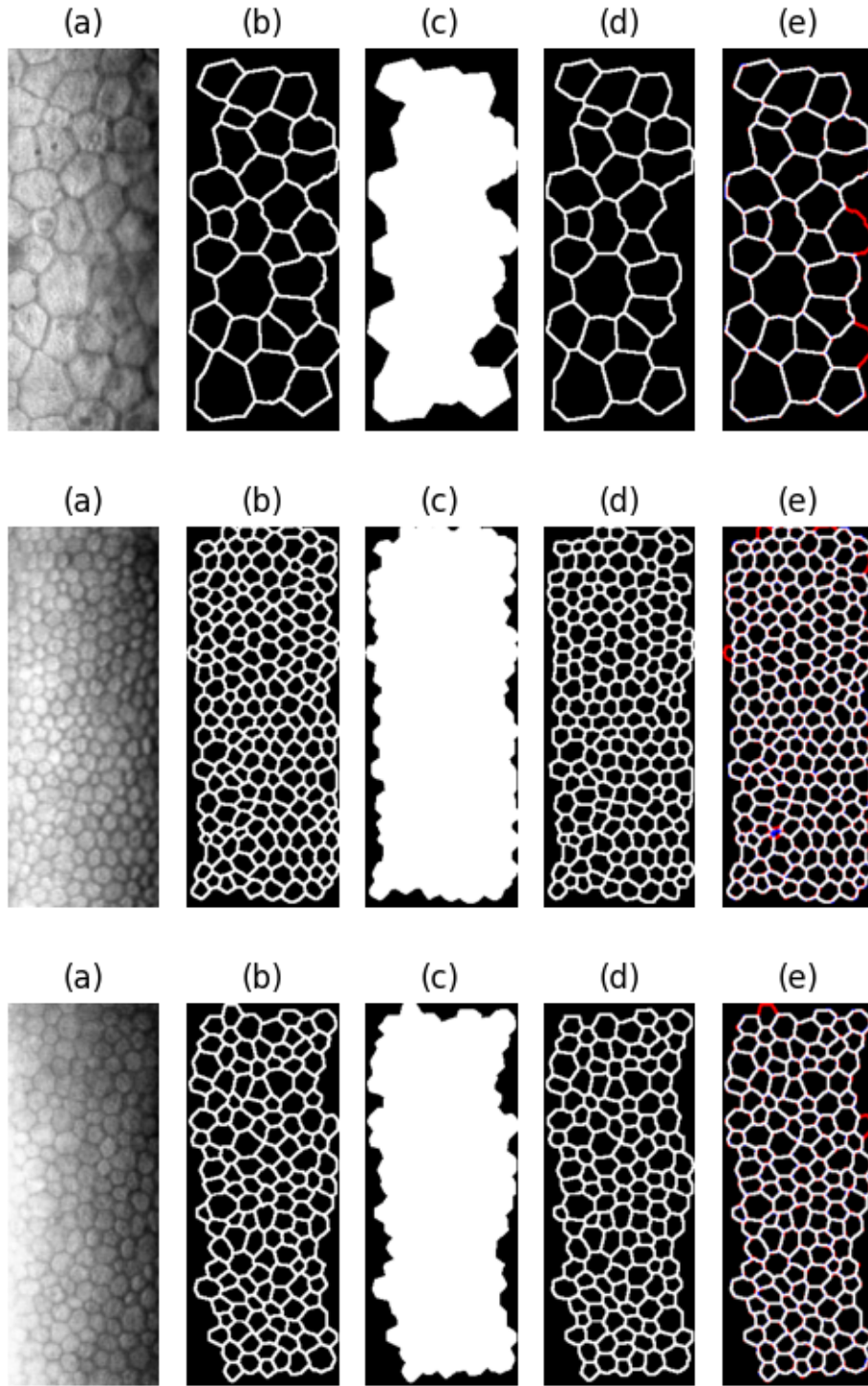


Figure 6: Dataset Gavet - (a) Original image. (b) Ground-truth segmentation mask. (c) Region of interest (ROI) extraction. (d) Final prediction after post-processing. (e) Visualisation of remaining segmentation errors.

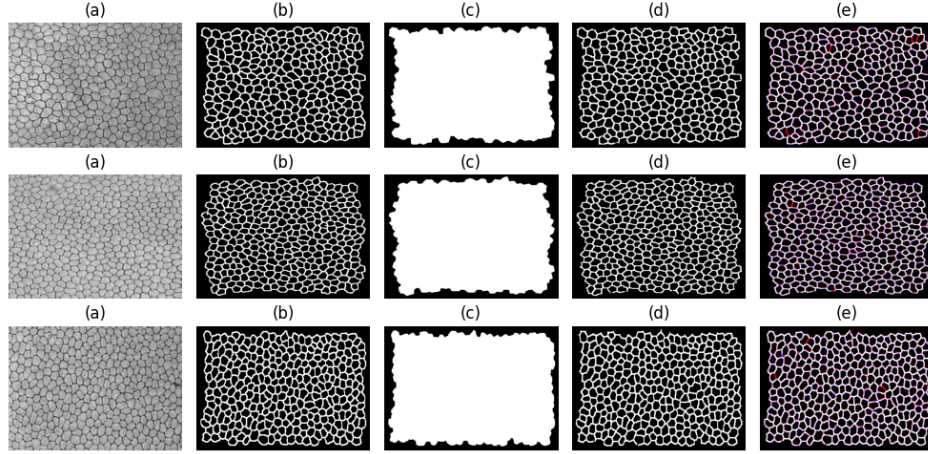


Figure 7: Dataset Alizarine - (a) Original image. (b) Ground-truth segmentation mask. (c) Region of interest (ROI) extraction. (d) Final prediction after post-processing. (e) Visualisation of remaining segmentation errors.

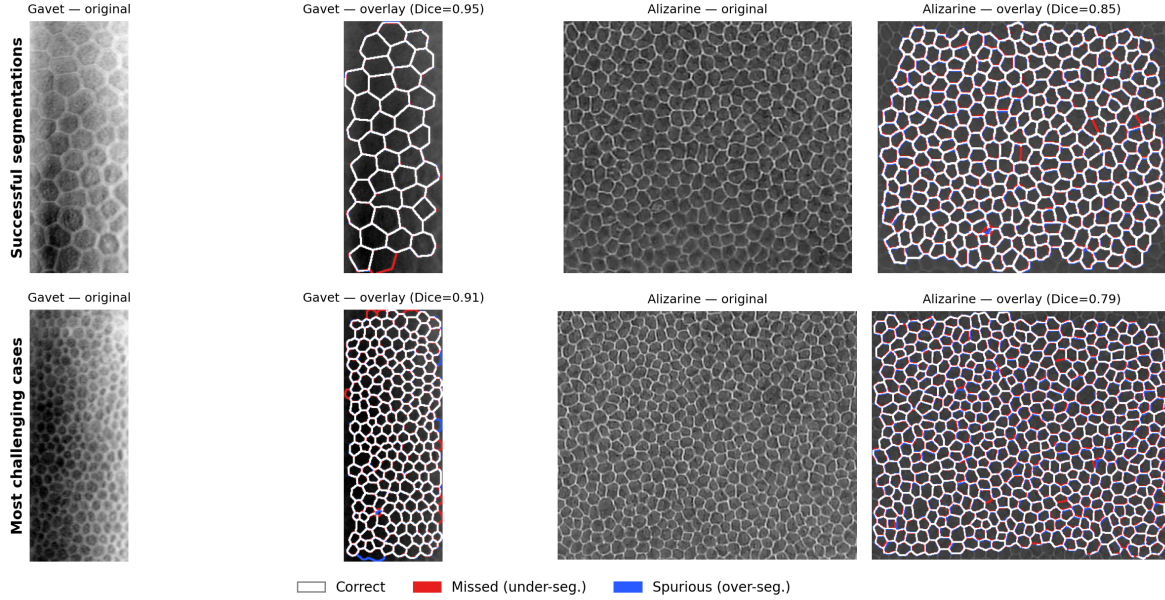


Figure 8: Representative success (top) and failure (bottom) cases of the proposed pipeline, shown as the best- and worst-Dice image of each dataset; for every case the original image is paired with the error overlay. *Top (successful)*: Gavet (Dice 0.95) and Alizarine (Dice 0.85), where the predicted mosaic closely matches the ground truth. *Bottom (most challenging)*: the worst Gavet image (Dice 0.91), whose errors are confined to the image periphery, and the worst Alizarine image (Dice 0.79), exhibiting under-segmentation by cell merging in low-contrast/out-of-focus areas and over-segmentation near stain artefacts. Overlay convention as in Figures 6–7: white = correct, red = missed (under-segmentation), blue = spurious (over-segmentation).

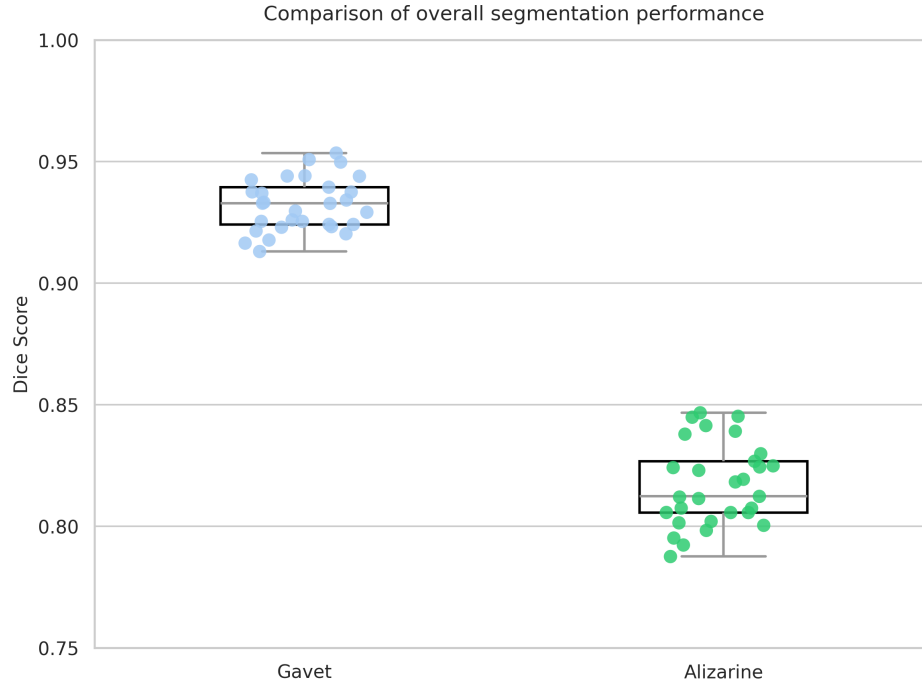


Figure 9: Comparison of DICE coefficients between the Gavet and Alizarine datasets. The horizontal line indicates the mean, and individual points show variability across images.

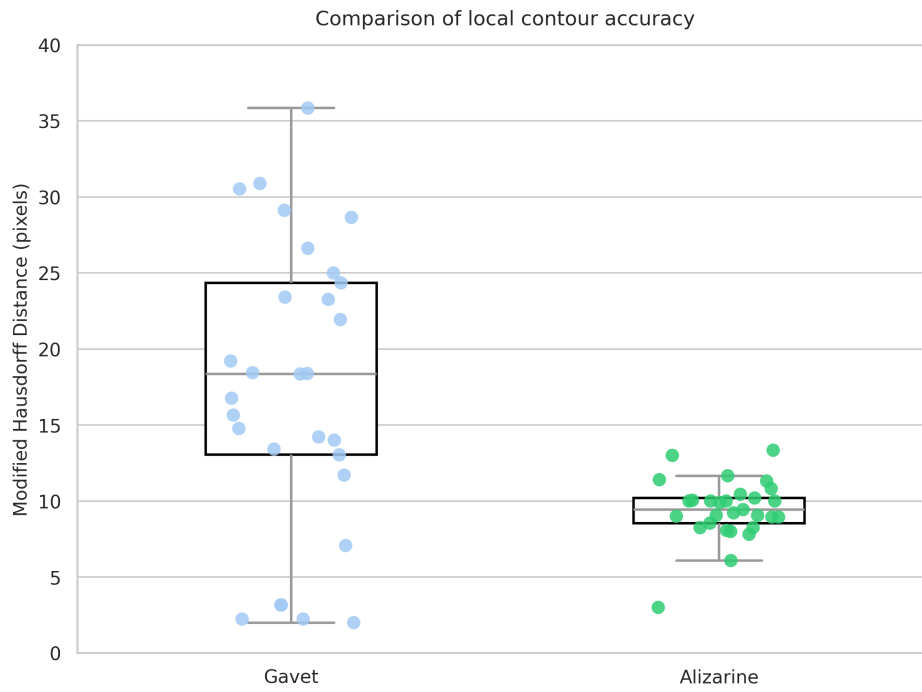


Figure 10: Comparison of the Modified Hausdorff Distance (MHD) between predicted and ground-truth contours for the Gavet and Alizarine datasets. Lower values indicate better local contour agreement.

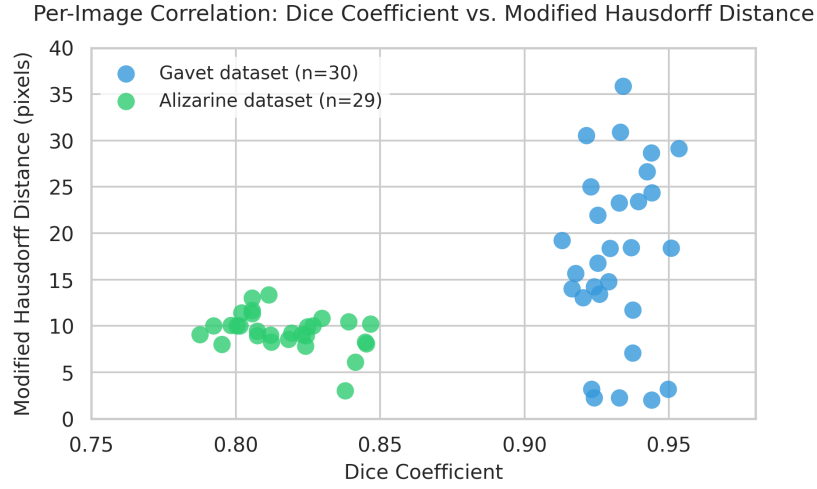


Figure 11: Per-image relationship between Dice and Modified Hausdorff Distance for the Gavet test set ( $n=30$ ) and the Alizarine dataset ( $n=29$ ). Mean Dice  $0.93 \pm 0.01$  on Gavet and  $0.82 \pm 0.02$  on Alizarine; the per-image MHD is tightly clustered on Alizarine ( $\approx 9$  px) and markedly more dispersed on Gavet, consistent with the higher MHD variability reported in Sec. 4. The two metrics are negatively correlated within each dataset but the correlation is not perfect — a few Alizarine images with intermediate Dice exhibit larger MHD, reflecting localized peripheral errors that have a limited impact on global overlap but a disproportionate effect on the worst-case localization error.

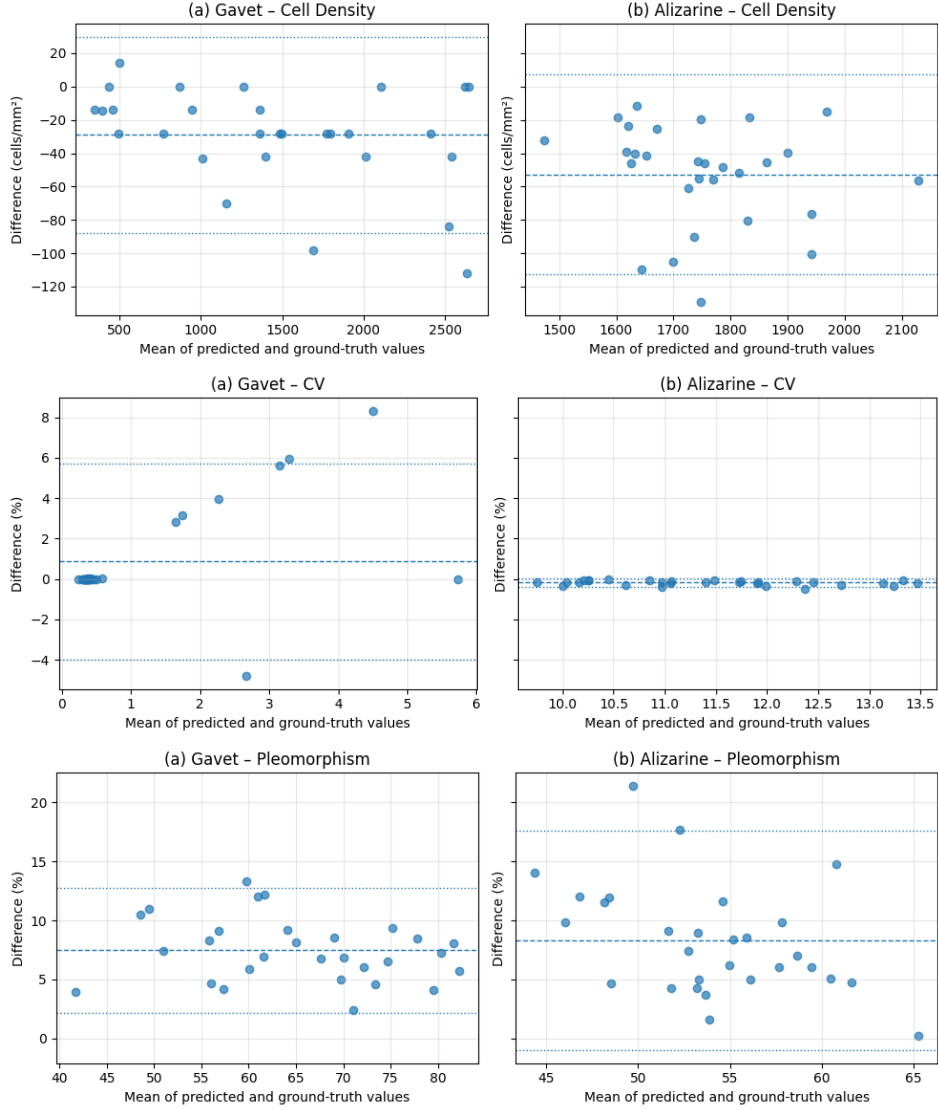


Figure 12: Bland–Altman plots comparing predicted and ground-truth morphometric parameters for the Gavet test set (top of each pair) and the Alizarine dataset (bottom of each pair). *Top*: ECD —  $x$ -axis: mean of predicted and reference ECD (cells/mm<sup>2</sup>);  $y$ -axis: difference predicted — reference (cells/mm<sup>2</sup>). *Middle*: CV —  $x$ -axis: mean CV (%);  $y$ -axis: difference predicted — reference (%). *Bottom*: Pleomorphism (HEX) —  $x$ -axis: mean HEX (%);  $y$ -axis: difference predicted — reference (%). In all panels the central dashed line indicates the mean difference (bias) and the dotted lines indicate the 95% limits of agreement (bias  $\pm 1.96\sigma$ ). Clinically, the absolute biases on ECD and CV remain well below the inter-observer variability of manual measurements [24, 3], while the larger HEX bias reflects the inherent sensitivity of shape-based parameters to local boundary errors.