Vesselpose: Vessel Graph Reconstruction from Learned Voxel-wise Direction Vectors in 3D Vascular Images

Rajalakshmi Palaniappan[1,2]　　　　　　　Rajalakshmi.Palaniappan@mdc-berlin.de
Christoph Karg[1]
Nemesio Navarro-Arambula[3]
Peter Hirsch[1,4] [iD]
Dagmar Kainmueller[*1,3,4] [iD]
Lisa Mais[*1,3,4] [iD]　　　　　　　　　　　Lisa.Mais@mdc-berlin.de
[1] *Max-Delbrueck-Center for Molecular Medicine in the Helmholtz Association (MDC),*
[2] *Humboldt University of Berlin,* [3] *University of Potsdam,* [4] *Helmholtz Imaging*

Abstract

Blood vessel segmentation and -tracing are essential tasks in many medical imaging applications. Although numerous methods exist, the prevailing segment-then-fix paradigm is fundamentally limited regarding its suitability for modelling the task of complete and topologically accurate vascular network reconstruction. We here propose an approach to extract topologically more accurate vascular graphs from 3D image data, building upon highly successful ideas from the related biomedical tasks of cell segmentation and -tracking. Our approach first predicts voxel-wise vessel direction vectors joint with standard vessel segmentation masks. Second, to extract the vascular graph from these predictions, we introduce a direction-vector-guided extension of the TEASAR algorithm. Our approach achieves state-of-the-art performance on three benchmark datasets, spanning both synthetic and real imagery. We further demonstrate the applicability of our approach to challenging 3D micro-CT scans of rat heart vasculature. Finally, we propose meaningful and interpretable measures of topological error, namely false splits and false merges for graphs. Overall, our approach substantially improves the topological accuracy of reconstructed vascular graphs, being able to separate closely apposed vessel segments and handle multiple vascular trees within a single volume.

Keywords: Blood Vessel Reconstruction, Centerline Topology, Evaluation.

1. Introduction

Tree-like structures are ubiquitous in living organisms and serve vital functions, for example as blood vessels, airways, or neuronal networks. Understanding how these branched systems develop, function, and change under pathological conditions requires detailed structural information. Image-based analysis has become a key tool for investigating such processes at the subcellular to organ scale, using diverse modalities such as MRI, micro-CT, electron microscopy, or light-sheet microscopy (Cheng et al., 2024; Walek et al., 2023; Damon-Soubeyrand et al., 2023; Todorov et al., 2020; Obenaus et al., 2017). To analyze these structures, image data are processed to extract simplified network representations—typically skeletons or graphs—from which features such as overall topology, segment
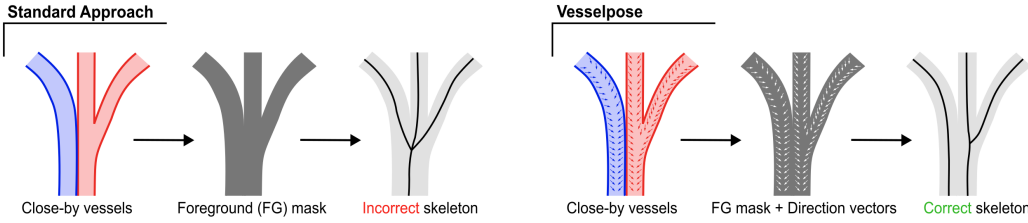
---

[*] Contributed equally

Figure 1: **Segmentation-and-skeletonization vs. Vesselpose.** Traditional segment-and-skeletonize pipelines often produce incorrect skeletons, especially when distinct vessels lie in close proximity. In contrast, Vesselpose leverages pixel-wise direction vectors to robustly reconstruct vascular trees, naturally handling closely apposed branches as well as multiple distinct trees.

lengths, and branching patterns can be quantified. For example, Liu et al. (2021) showed that altered topology of the microvasculature plays an important role in hepatocellular carcinoma (HCC). However, obtaining accurate and complete reconstructions remains challenging; manual or semi-automated tracing is still often the method of choice, particularly in cardiovascular imaging (Pampols-Perez et al., 2025; Rios Coronado et al., 2025).

Topological correctness refers to accurately preserving the connectivity of the biological network. Many vessel-graph extraction pipelines perform foreground-background segmentation of the image first (Tetteh et al., 2019; Todorov et al., 2020; Wittmann et al., 2025), and then skeletonize the foreground mask using TEASAR or variants of the Lee algorithm (Meyer-Spradow et al., 2009; Drees et al., 2021; Bumgarner and Nelson, 2022). However, these approaches struggle to achieve topological accuracy: Variable imaging contrast can render vessel segments faint or discontinuous, prompting segmentation methods to produce false splits; At the same time, branches running in close proximity often lead to false merges that incorrectly connect distinct vessels. In the subsequent skeletonization step, such false merge errors create artificial cycles or spurious branching points, as illustrated in the left part of Figure 1. Topological losses (Lux et al., 2025; Kirchhoff et al., 2024; Shit et al., 2021) or simple heuristics like thinning of ground truth masks may help with this issue to some extent – However, segmentation as a modular step remains fundamentally ill-suited for modelling the task of topologically correct vessel graph reconstruction. This limitation is not specific to foreground-background segmentation, but also holds for instance segmentation: Although the vasculature forms a globally connected system, imaging typically covers only a restricted anatomical region, yielding multiple disjoint trees; While instance segmentation can, in principle, model the separation of different vessel trees, it is not designed to prevent false merges *within* a single tree.

Therefore, alternative strategies are needed to achieve reliable and topologically accurate vascular graph reconstructions. Earlier work used combinatorial optimization to assemble vessel trees from small centerline tracklets (Türetken et al., 2010, 2011), yielding globally optimal tree reconstructions w.r.t. some objective under topological constraints, thereby ensuring topological correctness. However, even with heuristics and relaxed constraints, ILP-based methods (Türetken et al., 2016; Robben et al., 2014, 2016; Rempfler et al., 2016) remain computationally expensive and do not scale to large vascular networks. More recently, image-to-graph frameworks (Prabhakar et al., 2024; Naeem et al., 2024, 2025), inspired by DETR (Carion et al., 2020; Zhu et al., 2021), have emerged as a promising

2

direction. Most of them have been validated primarily on synthetic datasets, where Vesselformer (Prabhakar et al., 2024) still produces notable topological errors, while Trexplorer (Naeem et al., 2024) suffers from duplicate branching and premature tracking termination. Trexplorer-Super (Naeem et al., 2025) addresses these issues and extends evaluation to real datasets, yet its training and evaluation remain restricted to single-tree structures, whereas real vascular volumes typically contain multiple disjoint trees. Thus, a more general and computationally feasible solution is still required—one that can robustly extract topologically meaningful graphs from multi-tree vascular networks.

At the same time, issues of topological correctness have been addressed very successfully for the highly related tasks of cell segmentation and tracking in 3D(+t) microscopy data (Stringer et al., 2021; Malin-Mayor et al., 2023). Here, the community has moved away from the traditional segment-then-fix paradigm, replacing DL-based binary segmentation with models that predict pixel-wise shape properties that encode topologically relevant information (Hirsch and Kainmueller, 2020; Mais et al., 2020; Sheridan et al., 2023). Most prominently, Cellpose (Stringer et al., 2021; Pachitariu and Stringer, 2022) predicts vector fields pointing toward object centers; Similarly, for cell tracking through time, Malin-Mayor et al. (2023) predict pixel-wise direction vectors that point backward in time to the center of the same or mother cell in the previous frame. Iteratively following these vectors reconstructs complete cell lineages and ultimately traces each cell back to its origin. These advances highlight the value of predicting pixel-wise topological information, suggesting a promising direction that has not yet been extended to vascular tree reconstruction.

Building on these insights, we propose a method that extracts topologically plausible vessel trees from 3D images using a heuristic solver guided by pixel-wise predictions. We train a network to predict direction vectors that point toward the vessel centerline while being biased in the rootward direction, leveraging the anatomical prior that vessel diameter typically increases toward the root. This prior enables robust orientation along the tree and naturally suits vascular and airway networks. The predicted binary mask and direction vectors serve then as input to a novel skeletonization objective that reconstructs the tree structure by following the learned vector field. In summary, our main contributions are as follows:

- We present a DL-based method that predicts voxel-wise directional vectors from 3D vascular images, which a fast heuristic solver then assembles into a consistent vessel centerline graph.

- We introduce meaningful and easily interpretable topology-aware evaluation metrics such as false splits and false merges for graphs, proposing a tailor-made *assignment strategy* (cf. Maier-Hein et al. (2024)) based on hierarchical graph-matching.

- We outperform current SOTA on synthetic and real datasets and extend evaluation to a widely used multi-tree dataset (Tetteh et al., 2019) and a real 3D micro-CT dataset, achieving superior topological accuracy and reconstruction quality.

We make our code for the model and evaluation publicly available, along with trained models and prediction results, at the time of publication.
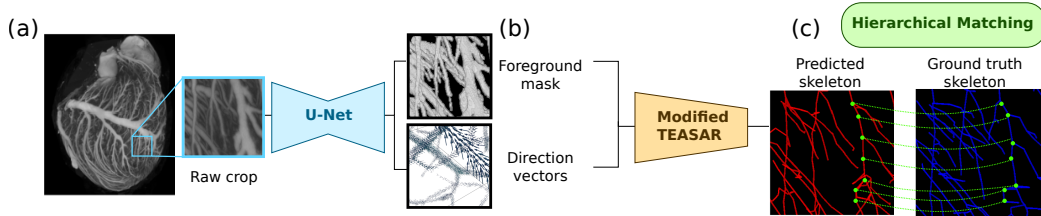
Figure 2: **Blood vessel reconstruction and evaluation.** (a) A U-Net predicts vessel foreground and voxel-wise direction vectors from the raw image. (b) A modified TEASAR algorithm extracts a skeleton graph. (c) Predicted skeletons are evaluated against ground truth using hierarchical graph matching as assignment strategy, which yields topologically meaningful error metrics.

2. Method

To derive graph representations from raw 3D vascular images, we first predict voxel-wise direction vectors capturing local structure, subsequently assembled into a tree-structured skeleton using a modified TEASAR algorithm (Sato et al., 2000). Figure 2 illustrates our approach. Section 2.1 describes direction vector generation and prediction, Section 2.2 the adaptation of TEASAR for vector-based skeletonization. Further postprocessing leverages direction vectors to handle falsely split components, as described in Suppl. Section A.3.

We formally represent vessel trees as directed acyclic graphs (Diestel, 2017): Nodes correspond to 3D coordinates marking either branching points or sampled points along vessel segments, while edges denote the segments connecting them. Each edge is assigned a radius characterizing the local vessel thickness. Edge directions follow a parent–child relationship, where the parent is the one closer to the vessel root.

2.1. Direction Vector Generation and Prediction

We train a 3D U-Net (Ronneberger et al., 2015) to jointly predict a foreground mask and voxel-wise direction vectors (x, y, z components) from 3D grayscale images. These vectors point toward the vessel centerline and are additionally biased rootward by a fixed stepsize, such that iterative following of the vectors eventually converges at the root. Given a ground truth foreground mask and a corresponding graph, we obtain training vectors as follows: For each foreground voxel, we first identify the nearest edge in the ground-truth graph within the local vessel radius. From the closest point on that edge, we then step a fixed distance toward the root, which defines the target point for the direction vector. The resulting vectors point from each voxel to this upstream centerline point, yielding smaller magnitudes for pixels close to the centerline and larger ones near the vessel boundary. Suppl. Figure 4 and Algorithm 1 provide details of the direction vectors and their generation. During training, we use Binary Cross-Entropy (BCE) loss for the foreground mask and Mean Squared Error (MSE) loss for the direction vectors.

2.2. TEASAR-based Centerline Generation

TEASAR (Sato et al., 2000) extracts tree-shaped skeletons from volumetric tubular segmentation masks by placing centerlines in regions that lie maximally far from the object

4

boundary. It implements this by iteratively tracing shortest paths from a root to any voxel within the mask, applying a penalty that discourages paths from approaching the boundary. This penalty depends solely on boundary distance, thus it may produce incorrect skeletons when vessels run in parallel, as shown in Figure 1. To address this, we propose a modified TEASAR variant that incorporates predicted voxel-wise direction vectors. These vectors exhibit minimal magnitude and smallest angle relative to the centerline direction at voxels closest to the centerline. We therefore augment the penalty term with components based on both *vector magnitude* and *angular deviation*. This additional penalty helps disambiguate vessels that are spatially tangent but semantically distinct.

Extended Penalty Term. Formally, let $\Omega \subset \mathbb{R}^3$ be the 3D object (foreground mask) and $\partial\Omega$ be the object boundary (vessel surface). Let $p \in \Omega$ be the voxel inside the object located at the end of the shortest path P, constructed in a previous iteration step. By $N \subseteq \Omega$ we denote the set of all those adjacent voxels of p are lying within the object. For each neighboring voxel $n \in N$, the *distance from boundary field (DBF)* of n, as leveraged by the original TEASAR, refers to the shortest Euclidean distance from voxel $n \in N$ to the nearest boundary point $b \in \partial\Omega$:

$$\text{DBF}(n) := \min_{b \in \partial\Omega} ||n - b|| \tag{1}$$

To incorporate directional guidance, we additionaly consider the *vector magnitude field (VMF)*, defined at each voxel n as the magnitude of its direction vector v_n: $\text{VMF}(n) := ||\boldsymbol{v}_n||$. Note that $\text{VMF}(n)$ is minimal if the voxel lies on the centerline. Furthermore we denote by $\theta(p, n) \in [0, 180]$ the angle (in degrees) between the direction vector $\boldsymbol{v}_p$ of p and the relative direction vector $\boldsymbol{r} := n - p$ from p to n. Again $\theta(p, n)$ is minimal if n is located in direction of the predicted direction vector $\boldsymbol{v}_p$. The adapted penalty value we propose is given by:

$$\text{PV}_{\text{flow}}(p, n) = 1{,}000{,}000 \cdot \left(\left(1 - \frac{\text{DBF}(n)}{M_1}\right)^{16} + \left(\frac{\text{VMF}(n)}{M_2}\right)^{16} + \left(\frac{\theta(p, n)}{M_3}\right)^{16} \right) \tag{2}$$

where

$$M_1 = \max_{p \in \Omega}(\text{DBF}(p))^{1.01}, \ \ M_2 = \max_{p \in \Omega}(\text{VMF}(p))^{1.01}, \ \ M_3 = 180 \tag{3}$$

This directly adopts the original TEASAR penalty, augmenting it with VMF- and θ- based terms of analogous form. Given this penalty, skeleton tracing proceeds as usual, starting from a most root-distant end point determined analogously as in original TEASAR, and appending the minimum-penalty neighboring node to the path until the root is reached.

Multi-root Processing and Adaptive Masking. Original TEASAR generates one skeleton per connected component of the binary mask. However, segmentation errors may merge distinct vessels into a single component, producing structures with multiple trees and therefore multiple roots. To address this, we extend TEASAR to support multiple roots within a component, as illustrated in Figure 3. Root locations in the datasets are provided either by manual annotation or automatically using the predicted direction vectors, effectively speeding up the annotation process. Details on automated root detection are provided in Suppl. Section A.1.
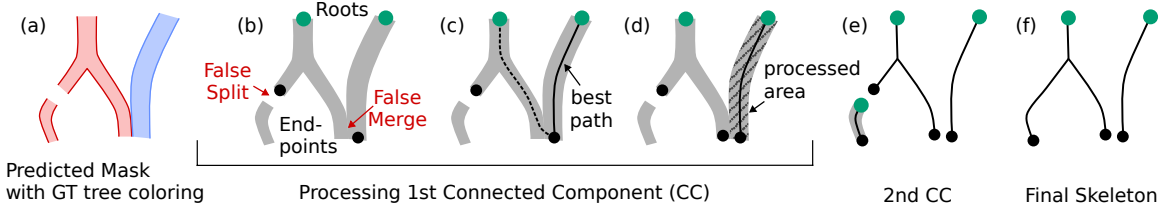
Figure 3: **Addressing topological errors with the modified TEASAR algorithm.**
(a) Predicted foreground mask with distinct ground truth (GT) trees having different colors. (b) The algorithm selects one connected component and identifies its roots and endpoints. (c) For each endpoint, paths are traced to all candidate roots, and the optimal path (with the lowest penalty) is chosen. (d) Voxels within a specified radius around the traced path are marked as processed and excluded from subsequent tracing. (e) After a component is fully processed, the algorithm proceeds to the next one; once all components are processed, disconnected fragments are evaluated for merging. (f) Final output with complete centerlines.

Once the best path is established (Figure 3(c) and (d)), original TEASAR applies a simple linear thresholding using a fixed scale and constant value $d = scale \cdot r + const$, where r is the vessel radius and d is the masking distance used to exclude already processed regions. We extend this with an adaptive masking scheme in which both parameters vary smoothly with the local vessel radius. This makes the method more robust across vessels of different radius and effectively suppresses spurious small branches in larger vessels. Further details are provided in Suppl. Section A.2.

3. Evaluation

Our method predicts an acyclic vessel skeleton, with 3D coordinates assigned to each node, and edges oriented towards the root, forming a labeled directed acyclic graph (DAG). Comparing such graph against respective ground-truth (GT) is challenging (Drees et al., 2019; Lyu et al., 2022): no standard metric exists (see Suppl. B.3), and many measures lack intuitive topological meaning or depend sensitively on node matching and sampling. To compare predicted and GT graphs, we first resample both at a fixed step size $s > 0$. The next essential step is an *assignment strategy* (cf. Maier-Hein et al. (2024)) that matches nodes and edges between predicted and GT graphs. In Sec. 3.1 we propose a greedy hierarchical matching procedure designed for robust topological correspondence. Based on these correspondences, we compute error metrics as described in Sec. 3.2.

3.1. Hierarchical Matching

Commonly used approaches to assign nodes or edges of two graphs to each other, is greedy nearest-neighbor- or optimal matching based solely on spatial proximity, such as in Drees et al. (2019); Naeem et al. (2025). While effective in simple scenarios, this strategy ignores the structural and semantic information inherent to tree-like graphs, making it unsuitable for capturing topological similarity—particularly in cases where different vessels are close-by. Therefore, we propose a greedy one-to-at-most-one hierarchical matching scheme that

incorporates spatial, semantic and ancestor information. It is similar to Gillette et al. (2011), but also applicable in multi tree scenarios.

A pseudo-code description of the matching procedure is provided in Suppl. Algorithm 2. In short, first, the connected components of the GT graph G and predicted graph P are determined. Each node in G and P is assigned a semantic class—*root*, *branching point*, *leaf*, or *intermediate*. For every node in G, we identify candidate nearest neighbors in P within a predefined distance threshold and rank them, first by semantic correspondence and second by spatial proximity. We then iterate over the GT roots, always selecting next the root whose best candidate exhibits the highest matching priority (i.e., first by identical semantic class and second by minimal distance). Starting from each root, we perform *two depth-first traversals*. In the first, we visit branching and leaf nodes, and assign them to the best available candidate based on the matching status of the candidate's parent, the candidate's semantic label, and its distance. Thus, candidates whose parents are matched within the same GT tree receive highest priority. If no suitable candidate exists, the GT node remains unmatched. The second traversal processes intermediate nodes using the same criteria. After completing a GT tree, we proceed to the next. Importantly, the candidate lists are updated immediately whenever two nodes become matched to maintain consistency throughout the hierarchy. A quantitative comparison with greedy nearest-neighbor and Hungarian matching is presented in Suppl. Table 3.

3.2. Metrics Definitions

Based on our literature review in Suppl. Section B.3, we report the edge-wise F1 score—a widely established and, as we find, easily interpretable measure of topological correctness. Yet F1 alone does not capture the structural impact of certain errors. For instance, a false positive edge connecting unrelated nodes can distort the topology far more than a shortcut to an ancestor (see Suppl. Figure 5). To address this, we introduce *false splits* and *false merges* as additional topology-aware error measures, extending prior work (Matula et al., 2015; Mais et al., 2024) to multi-tree graphs where errors may arise both within and across trees. In addition, we report the metrics used in Naeem et al. (2025) for comparability reasons. Although they also report F1 scores at the node and branch level—similar in spirit to our recommendation—their metrics rely on greedy one-to-one nearest-neighbor matching and the computation operates on individual nodes, thereby not fully capturing connectivity. Regarding graph-level Betti numbers, only Betti–0 (the number of connected components) is meaningful; because assuming only trees, Betti–1 (the number of cycles) is always zero.

In the following, we denote by G and P the GT and predicted graphs with node sets V_G and V_P, and by $\Phi : V_G \to V_P$ the one-to-one node matching. Both graphs are resampled to a fixed step size ($s = 1$ voxel, unless stated otherwise), after which we apply our proposed hierarchical matching. The **edge-wise F1 score** is computed as the balanced measure of precision and recall, relating the number of correctly matched edges (TP) in G to the number of incorrectly matched (FP) or incorrectly unmatched (FN) edges. We additionally report edge-wise precision and recall. A formal definition is given in Suppl. Section B.2.

We further define **false merges (FM)** and **false splits (FS)** as follows. A false merge is a FP edge $(\Phi(v), \Phi(v'))$ in P where the GT nodes v and v' in G have no directed path between them. In other words, neither is an ancestor of the other. For a false split, we

consider the subgraph P^* of P obtained by excluding all false merge edges. A false split is a FN edge (v, v') in G such that adding the missing edge $(\Phi(v), \Phi(v'))$ to P^* merges two connected components into one. This means FS correspond to missing edges in P that cause false disconnections in P^*. The number of FS can be determined by $\beta_0(P^*) - \beta_0(P)$, since each FS increases the number of connected components of P^*. Here, β_0 is the number of connected components (Betti-0).

To compute the metrics from Naeem et al. (2025), both G and P are resampled to one voxel spacing. At the node level, precision, recall, and F1 are reported, along with radius accuracy measured via the mean absolute error (MAE). At the branch level, the F1 score is reported; where a branch is considered a TP if at least 80% of its nodes are matched.

4. Experiments

We evaluate our method on four vascular datasets, covering both single and multi-tree scenarios. Consistent with prior works (Naeem et al., 2025, 2024; Prabhakar et al., 2024), all reported experiments use ground-truth root locations as input to our adapted TEASAR algorithm. Dataset-specific modifications of the training procedure, together with details on sample sizes and the data splits, are in Suppl. Section C. An ablation study quantifying the contributions of each components of our method is included in Suppl. Section C.5.

4.1. Model Architecture and Training

For segmentation and vector prediction, we employ a 4-layer U-Net with batch normalization and 16 initial feature channels, which double at each downsampling step. The network is trained on randomly sampled input patches. Data augmentation includes intensity shifts and randomly masking out $3 \times 3 \times 3$ crops. Training is performed for 400,000 iterations with a batch size of 1 using the Adam optimizer. Aside from dataset-specific input sizes and augmentations, the architecture and training protocol are kept identical across all datasets.

4.2. Case 1: Single-Tree Data

For the single-tree datasets, we compare our method against Vesselformer (Prabhakar et al., 2024), Trexplorer (Naeem et al., 2024), and Trexplorer-super (Naeem et al., 2025). We report all baselines and metrics as in Naeem et al. (2025), as we were unable to reproduce their published results and thus cannot faithfully compare in terms of our new metrics. Instead, we follow their evaluation protocol to enable a fair comparison. Thus we report point-level F1, precision, recall, and radius MAE, as well as branch-level F1 and Betti scores.

The **Single-Tree Synthetic** dataset, introduced in Naeem et al. (2025), is generated using the Synthetic Vascular Toolkit (SVT) (Sexton et al., 2025). Each volume contains a single vascular tree, its segmentation mask, and the corresponding 3D centerline graph. As shown in Table 1, our model consistently outperforms the current state of the art across both point-level and branch-level metrics. Suppl. Figure 6 shows qualitative results.

The publicly available **Parse 2022** pulmonary artery segmentation dataset (Luo et al., 2024) contains 100 computed tomography pulmonary angiography (CTPA) volumes with pixel-wise segmentation masks. These masks were created semi-automatically by experts using a region-growing approach. Naeem et al. (2025) subsequently derived centerlines from

Table 1: Quantitative comparison of our method with Vesselformer, Trexplorer and Trexplorer Super for the Single-Tree Synthetic and Parse 2022 datasets. Please note that we report all baselines and metrics as in Naeem et al. (2025).

Model	Point Level				Branch Level	Graph Level	
	F1↑	Prec↑	Rec↑	Rad.(MAE)↓	F1↑	$\beta_0 \downarrow$	$\beta_1 \downarrow$
Synthetic Vesselformer	48.18	44.53	61.52	0.42	15.95	81.7	653.5
Trexplorer	39.40	30.91	78.21	0.23	26.26	0	0
Trexpl. Super	77.83	91.91	70.44	**0.1**	77.12	0	0
Ours	**90.41**	**92.32**	**88.60**	0.37	**80.71**	0	0
Parse2022 Vesselformer	16.43	18.49	15.28	1.11	1.99	410	246.7
Trexplorer	10.01	9.87	12.01	1.21	3.71	0	0
Trexpl. Super	39.46	**55.27**	33.99	**0.56**	23.46	0	0
Ours	**49.42**	43.85	**59.21**	0.588	**28.87**	2.70	0

these masks using the Kimimaro TEASAR implementation (Silversmith et al., 2021). Note that these ground-truth centerlines were generated *automatically*. As shown in Table 1, our model outperforms the current state of the art on both point-level and branch-level F1. We observe lower point-level precision, which we attribute to the model predicting small vessel structures that are not annotated in the ground truth masks (see Suppl. Figure 7). We also note some Betti-0 errors: although our post-processing step is designed to correct false splits, it does not fully guarantee global connectivity of the predicted vascular tree.

4.3. Case 2: Multi-Tree Data

Although vascular networks are ideally single-tree structures, real data often contain multiple trees due to challenges in separating arteries and veins or imaging artifacts. Here, we report our recommended metrics, namely edge-level F1, precision, recall, false merges (FM), and false splits (FS). We compare Vesselpose against different segmentation-based approaches, where we skeletonize the resulting binary masks with Kimimaro TEASAR.

The **Multi-Tree Synthetic** dataset originates from Tetteh et al. (2019) and is generated using vessel formation simulations (Schneider et al., 2012). We compare our method to vesselFM (Wittmann et al., 2025) and a standard U-Net (Ronneberger et al., 2015). We also report an upper bound by applying TEASAR directly to the ground truth masks of Schneider et al. (2012). The results in Table 2 show that our method consistently outperforms these segmentation-based baselines. Original TEASAR produces one tree per connected component, but baseline segmentations frequently merge distinct trees into a single component. As a result, their false merge rates are substantially higher than our method. In Suppl. Figure 8 we show qualitative results and discuss failure cases of our method.

The **Multi-Tree Micro-CT** data were acquired from perfused rat hearts using a solidifying Microfil agent (Napieczyńska et al., 2024), which provides strong contrast and reveals even small vessels. The dataset is still under study and may be made publicly available at a later stage. We use four rat heart volumes: one to fine-tune a U-Net pretrained on

Table 2: Quantitative comparison of our method on the Multi-Tree Synthetic and Micro-CT Heart datasets. We compare against U-Net, vesselFM, and the ground truth segmentation, each followed by skeletonization. Because a more complete prediction can yield higher absolute FM and FS counts than an incomplete graph, we additionally report relative FM/FS values, obtained by dividing the absolute counts by the total number of predicted edges.

	Method	Edges			FM↓		FS↓	
		F1↑	Prec↑	Rec↑	Rel.	Abs.	Rel.	Abs.
Multi-Tree Synthetic	U-Net+TEASAR	0.61	0.72	0.52	0.02	51.2	0.01	52.9
	VesselFM+TEASAR	0.60	0.71	0.53	0.02	54.4	0.01	61.4
	GT Segm.+TEASAR	0.61	0.72	0.54	0.02	53.8	0.009	40.1
	Ours	**0.80**	**0.80**	**0.79**	**0.007**	**29.7**	**0.006**	**28.3**
Micro CT	U-Net+TEASAR	0.37	0.45	0.31	0.01	**56**	0.01	**51**
	Ours	**0.50**	**0.59**	**0.43**	**0.008**	70.5	**0.008**	71.5

the synthetic multi-tree data, and three for validation and testing. For these, we annotated three $400 \times 400 \times 400$ crops using CATMAID (Saalfeld et al., 2009; Schneider-Mizell et al., 2016). Details on the data, model, and fine-tuning procedure are provided in Suppl. Section C.4. Table 2 shows that our method consistently outperforms a standard U-Net with TEASAR. Our higher absolute FM and FS values stem from reconstructing more complete skeletons, whereas U-Net and regular TEASAR miss large graph regions—reflected in our correspondingly lower relative FM/FS counts, also seen in Suppl. Figure 9.

5. Conclusion

In this work, we presented a novel method for extracting vessel graphs from 3D images. We demonstrated its effectiveness on four datasets, comprising synthetic and real data, as well as single-tree and multi-tree vascular structures. In addition, we introduced a hierarchical graph matching algorithm that yields more topologically meaningful node and edge correspondences, and we defined false splits and false merges as intuitive topology-aware error measures for tree graphs. Despite these advances, several evaluation metrics such as the edge-wise F1 score remain sensitive to different node sampling and matching strategies. A more systematic analysis of these effects represents a valuable direction for future work. Moreover, there is a pressing need for publicly available, real-world datasets with high-quality, manually annotated 3D vessel graphs. Such datasets, combined with an established evaluation protocol encompassing sampling, matching, and metrics, would be essential for enabling consistent benchmarking and further methodological development in this area.

References

Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, Oct 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

Jacob R. Bumgarner and Randy J. Nelson. Open-source analysis and visualization of segmented vasculature datasets with vesselvio. *Cell Reports Methods*, 2(4):100189, 2022. ISSN 2667-2375. doi: https://doi.org/10.1016/j.crmeth.2022.100189. URL https://www.sciencedirect.com/science/article/pii/S2667237522000443.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 213–229, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58452-8.

Siyuan Cheng, Ivan Fan Xia, Renate Wanner, Javier Abello, Amber N Stratman, and Stefania Nicoli. Hemodynamics regulate spatiotemporal artery muscularization in the developing circle of willis. *eLife*, 13:RP94094, jul 2024. ISSN 2050-084X. doi: 10.7554/eLife.94094. URL https://doi.org/10.7554/eLife.94094.

Christelle Damon-Soubeyrand, Antonino Bongiovanni, Areski Chorfa, Chantal Goubely, Nelly Pirot, Luc Pardanaud, Laurence Piboin-Fragner, Caroline Vachias, Stephanie Bravard, Rachel Guiton, Jean-Leon Thomas, Fabrice Saez, Ayhan Kocer, Meryem Tardivel, Joël R Drevet, and Joelle Henry-Berger. Three-dimensional imaging of vascular development in the mouse epididymis. *eLife*, 12:e82748, jun 2023. ISSN 2050-084X. doi: 10.7554/eLife.82748. URL https://doi.org/10.7554/eLife.82748.

Reinhard Diestel. *Graph Theory*. Graduate Texts in Mathematics. Springer, 5 edition, 2017. ISBN 978-3-319-41320-4.

Dominik Drees, Aaron Scherzinger, and Xiaoyi Jiang. Gerome-a method for evaluating stability of graph extraction algorithms without ground truth. *IEEE Access*, 7:21744–21755, 2019. doi: 10.1109/ACCESS.2019.2898754.

Dominik Drees, Aaron Scherzinger, René Hägerling, Friedemann Kiefer, and Xiaoyi Jiang. Scalable robust graph and feature extraction for arbitrary vessel networks in large volumetric datasets. *BMC Bioinformatics*, 22(1):346, Jun 2021. ISSN 1471-2105. doi: 10.1186/s12859-021-04262-w. URL https://doi.org/10.1186/s12859-021-04262-w.

Adrien Foucart, Olivier Debeir, and Christine Decaestecker. Panoptic quality should be avoided as a metric for assessing cell nuclei segmentation and classification in digital pathology. *Scientific Reports*, 13(1):8614, May 2023. ISSN 2045-2322. doi: 10.1038/s41598-023-35605-7. URL https://doi.org/10.1038/s41598-023-35605-7.

Alejandro Frangi, W.J. Niessen, Koen Vincken, and Max Viergever. Multiscale vessel enhancement filtering. *Med. Image Comput. Comput. Assist. Interv.*, 1496, 02 2000.

Todd A. Gillette, Keith M. Brown, and Giorgio A. Ascoli. The diadem metric: Comparing multiple reconstructions of the same neuron. *Neuroinformatics*, 9(2-3):233–245, 2011. doi: 10.1007/s12021-011-9117-y.

Peter Hirsch and Dagmar Kainmueller. An auxiliary task for learning nuclei segmentation in 3d microscopy images. In *International Conference on Medical Imaging with Deep Learning, MIDL 2020, 6-8 July 2020, Montréal, QC, Canada*, volume 121 of *Proceedings of Machine Learning Research*, pages 304–321. PMLR, 2020. URL http://proceedings.mlr.press/v121/hirsch20a.html.

Yannick Kirchhoff, Maximilian R. Rokuss, Saikat Roy, Balint Kovacs, Constantin Ulrich, Tassilo Wald, Maximilian Zenk, Philipp Vollmuth, Jens Kleesiek, Fabian Isensee, and Klaus Maier-Hein. Skeleton recall loss for connectivity conserving and resource efficient segmentation of thin tubular structures. In Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol, editors, *Computer Vision – ECCV 2024*, pages 218–234, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72980-5.

Wenjian Li, Guannan Wu, and Huiqi Li. Similarity evaluation of retinal vascular network based on tree edit distance. In *2023 IEEE 18th Conference on Industrial Electronics and Applications (ICIEA)*, pages 722–727, 2023. doi: 10.1109/ICIEA58696.2023.10241828.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

Qiaoyu Liu, Boyu Zhang, Luna Wang, Rencheng Zheng, Jinwei Qiang, He Wang, Fuhua Yan, and Ruokun Li. Assessment of vascular network connectivity of hepatocellular carcinoma using graph-based approach. *Frontiers in Oncology*, 11:668874, 2021.

Gongning Luo, Kuanquan Wang, Jun Liu, Shuo Li, Xinjie Liang, Xiangyu Li, Shaowei Gan, Wei Wang, Suyu Dong, Wenyi Wang, Pengxin Yu, Enyou Liu, Hongrong Wei, Na Wang, Jia Guo, Huiqi Li, Zhao Zhang, Ziwei Zhao, Na Gao, Nan An, Ashkan Pakzad, Bojidar Rangelov, Jiaqi Dou, Song Tian, Zeyu Liu, Yi Wang, Ampatishan Sivalingam, Kumaradevan Punithakumar, Zhaowen Qiu, and Xin Gao. Efficient automatic segmentation for multi-level pulmonary arteries: The parse challenge, 2024. URL https://arxiv.org/abs/2304.03708.

Laurin Lux, Alexander H Berger, Alexander Weers, Nico Stucki, Daniel Rueckert, Ulrich Bauer, and Johannes C. Paetzold. Topograph: An efficient graph-based framework for strictly topology preserving image segmentation. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=QOzmmNNePz.

Xingzheng Lyu, Li Cheng, and Sanyuan Zhang. The reta benchmark for retinal vascular tree analysis. *Scientific Data*, 9(1):397, Jul 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01507-y. URL https://doi.org/10.1038/s41597-022-01507-y.

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, A. Emre Kavur, Carole H. Sudre, Michael Baumgartner, Matthias Eisenmann, Doreen Heckmann-Nötzel, Tim Rädsch, Laura Acion, Michela Antonelli, Tal Arbel, Spyridon Bakas, Arriel Benis, Matthew B. Blaschko, M. Jorge Cardoso, Veronika Cheplygina, Beth A. Cimini, Gary S. Collins, Keyvan Farahani, Luciana Ferrer, Adrian Galdran, Bram van Ginneken, Robert Haase, Daniel A. Hashimoto, Michael M. Hoffman, Merel Huisman, Pierre Jannin, Charles E. Kahn, Dagmar Kainmueller, Bernhard Kainz, Alexandros Karargyris, Alan Karthikesalingam, Florian Kofler, Annette Kopp-Schneider, Anna Kreshuk, Tahsin Kurc, Bennett A. Landman, Geert Litjens, Amin Madani, Klaus Maier-Hein, Anne L. Martel, Peter Mattson, Erik Meijering, Bjoern Menze, Karel G. M. Moons, Henning Müller, Brennan Nichyporuk, Felix Nickel, Jens Petersen, Nasir Rajpoot, Nicola Rieke, Julio Saez-Rodriguez, Clara I. Sánchez, Shravya Shetty, Maarten van Smeden, Ronald M. Summers, Abdel A. Taha, Aleksei Tiulpin, Sotirios A. Tsaftaris, Ben Van Calster, Gaël Varoquaux, and Paul F. Jäger. Metrics reloaded: recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, Feb 2024. ISSN 1548-7105. doi: 10.1038/s41592-023-02151-z. URL https://doi.org/10.1038/s41592-023-02151-z.

Lisa Mais, Peter Hirsch, and Dagmar Kainmueller. Patchperpix for instance segmentation. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 288–304, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58595-2.

Lisa Mais, Peter Hirsch, Claire Managan, Ramya Kandarpa, Josef Lorenz Rumberger, Annika Reinke, Lena Maier-Hein, Gudrun Ihrke, and Dagmar Kainmueller. Fisbe: A real-world benchmark dataset for instance segmentation of long-range thin filamentous structures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22249–22259, June 2024.

Caroline Malin-Mayor, Peter Hirsch, Leo Guignard, Katie McDole, Yinan Wan, William C. Lemon, Dagmar Kainmueller, Philipp J. Keller, Stephan Preibisch, and Jan Funke. Automated reconstruction of whole-embryo cell lineages by learning from sparse annotations. *Nature Biotechnology*, 41(1):44–49, Jan 2023. ISSN 1546-1696. doi: 10.1038/s41587-022-01427-7. URL https://doi.org/10.1038/s41587-022-01427-7.

Pavel Matula, Martin Maška, Dmitry V. Sorokin, Petr Matula, Carlos Ortiz-de Solórzano, and Michal Kozubek. Cell tracking accuracy measurement based on comparison of acyclic oriented graphs. *PLOS ONE*, 10(12):1–19, 12 2015. doi: 10.1371/journal.pone.0144959. URL https://doi.org/10.1371/journal.pone.0144959.

Jennis Meyer-Spradow, Timo Ropinski, Jörg Mensmann, and Klaus H. Hinrichs. Voreen: A rapid-prototyping environment for ray-casting-based volume visualizations. *IEEE Computer Graphics and Applications*, 29:6–13, 2009. URL https://api.semanticscholar.org/CorpusID:8211514.

Roman Naeem, David Hagerman, Lennart Svensson, and Fredrik Kahl. Trexplorer: Recurrent DETR for Topologically Correct Tree Centerline Tracking . In *proceedings of Medical*

Image Computing and Computer Assisted Intervention – MICCAI 2024, volume LNCS 15011. Springer Nature Switzerland, October 2024.

Roman Naeem, David Hagerman, Jennifer Alvén, Lennart Svensson, and Fredrik Kahl. Trexplorer super: Topologically correct centerline tree tracking of tubular objects in ct volumes. In James C. Gee, Daniel C. Alexander, Jaesung Hong, Juan Eugenio Iglesias, Carole H. Sudre, Archana Venkataraman, Polina Golland, Jong Hyo Kim, and Jinah Park, editors, *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pages 595–605, Cham, 2025. Springer Nature Switzerland. ISBN 978-3-032-04984-1.

Hanna Napieczyńska, Sarah M Kedziora, Nadine Haase, Dominik N Müller, Arnd Heuser, Ralf Dechend, and Kristin Kräker. μct imaging of a multi-organ vascular fingerprint in rats. *Plos one*, 19(10):e0308601, 2024.

Andre Obenaus, Michelle Ng, Amanda M. Orantes, Eli Kinney-Lang, Faisal Rashid, Mary Hamer, Richard A. DeFazio, Jiping Tang, John H. Zhang, and William J. Pearce. Traumatic brain injury results in acute rarefication of the vascular network. *Scientific Reports*, 7(1):239, Mar 2017. ISSN 2045-2322. doi: 10.1038/s41598-017-00161-4. URL https://doi.org/10.1038/s41598-017-00161-4.

Marius Pachitariu and Carsen Stringer. Cellpose 2.0: how to train your own model. *Nature Methods*, 19(12):1634–1641, Dec 2022. ISSN 1548-7105. doi: 10.1038/s41592-022-01663-4. URL https://doi.org/10.1038/s41592-022-01663-4.

Mireia Pampols-Perez, Carina Fürst, Oscar Sánchez-Carranza, Elena Cano, Jonathan Alexis Garcia-Contreras, Lisa Mais, Wenhan Luo, Sandra Raimundo, Eric L. Lindberg, Martin Taube, Arnd Heuser, Anje Sporbert, Dagmar Kainmueller, Miguel O. Bernabeu, Norbert Hübner, Holger Gerhardt, Gary R. Lewin, and Annette Hammes. Mechanosensitive piezo2 channels shape coronary artery development. *Nature Cardiovascular Research*, 4 (7):921–937, Jul 2025. ISSN 2731-0590. doi: 10.1038/s44161-025-00677-3. URL https://doi.org/10.1038/s44161-025-00677-3.

Chinmay Prabhakar, Suprosanna Shit, Johannes C. Paetzold, Ivan Ezhov, Rajat Koner, Hongwei Li, Florian Sebastian Kofler, and Bjoern Menze. Vesselformer: Towards complete 3d vessel graph generation from images. In Ipek Oguz, Jack Noble, Xiaoxiao Li, Martin Styner, Christian Baumgartner, Mirabela Rusu, Tobias Heinmann, Despina Kontos, Bennett Landman, and Benoit Dawant, editors, *Medical Imaging with Deep Learning*, volume 227 of *Proceedings of Machine Learning Research*, pages 320–331. PMLR, 10–12 Jul 2024. URL https://proceedings.mlr.press/v227/prabhakar24a.html.

Markus Rempfler, Bjoern Andres, and Bjoern H Menze. The minimum cost connected subgraph problem in medical image analysis. In *Medical Image Computing and Computer-Assisted Intervention-MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III 19*, pages 397–405. Springer, 2016.

Pamela E. Rios Coronado, Jiayan Zhou, Xiaochen Fan, Daniela Zanetti, Jeffrey A. Naftaly, Pratima Prabala, Azalia M. Martínez Jaimes, Elie N. Farah, Soumya Kundu, Salil S. Deshpande, Ivy Evergreen, Pik Fang Kho, Qixuan Ma, Austin T. Hilliard,

Sarah Abramowitz, Saiju Pyarajan, Daniel Dochtermann, Scott M. Damrauer, Kyong-Mi Chang, Michael G. Levin, Virginia D. Winn, Anca M. Paşca, Mary E. Plomondon, Stephen W. Waldo, Philip S. Tsao, Anshul Kundaje, Neil C. Chi, Shoa L. Clarke, Kristy Red-Horse, and Themistocles L. Assimes. ¡em¿cxcl12¡/em¿ drives natural variation in coronary artery anatomy across diverse populations. *Cell*, 188(7): 1784–1806.e22, Apr 2025. ISSN 0092-8674. doi: 10.1016/j.cell.2025.02.005. URL https://doi.org/10.1016/j.cell.2025.02.005.

David Robben, Engin Türetken, Stefan Sunaert, Vincent Thijs, Guy Wilms, Pascal Fua, Frederik Maes, and Paul Suetens. Simultaneous segmentation and anatomical labeling of the cerebral vasculature. In Polina Golland, Nobuhiko Hata, Christian Barillot, Joachim Hornegger, and Robert Howe, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*, pages 307–314, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10404-1.

David Robben, Engin Türetken, Stefan Sunaert, Vincent Thijs, Guy Wilms, Pascal Fua, Frederik Maes, and Paul Suetens. Simultaneous segmentation and anatomical labeling of the cerebral vasculature. *Medical Image Analysis*, 32:201–215, 2016. ISSN 1361-8415. doi: https://doi.org/10.1016/j.media.2016.03.006. URL https://www.sciencedirect.com/science/article/pii/S1361841516300056.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In Nassir Navab, Joachim Hornegger, William M. Wells, and Alejandro F. Frangi, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pages 234–241, Cham, 2015. Springer International Publishing. ISBN 978-3-319-24574-4.

Stephan Saalfeld, Albert Cardona, Volker Hartenstein, and Pavel Tomančák. Catmaid: collaborative annotation toolkit for massive amounts of image data. *Bioinformatics*, 25 (15):1984–1986, 04 2009. ISSN 1367-4803. doi: 10.1093/bioinformatics/btp266. URL https://doi.org/10.1093/bioinformatics/btp266.

M. Sato, I. Bitter, M.A. Bender, A.E. Kaufman, and M. Nakajima. Teasar: tree-structure extraction algorithm for accurate and robust skeletons. In *Proceedings the Eighth Pacific Conference on Computer Graphics and Applications*, pages 281–449, 2000. doi: 10.1109/PCCGA.2000.883951.

Matthias Schneider, Johannes Reichold, Bruno Weber, Gábor Székely, and Sven Hirsch. Tissue metabolism driven arterial tree generation. *Medical image analysis*, 16(7):1397–1414, 2012.

Casey M Schneider-Mizell, Stephan Gerhard, Mark Longair, Tom Kazimiers, Feng Li, Maarten F Zwart, Andrew Champion, Frank M Midgley, Richard D Fetter, Stephan Saalfeld, and Albert Cardona. Quantitative neuroanatomy for connectomics in *Drosophila*. *eLife*, 5:e12059, mar 2016. ISSN 2050-084X. doi: 10.7554/eLife.12059. URL https://doi.org/10.7554/eLife.12059.

Zachary A. Sexton, Dominic Rütsche, Jessica E. Herrmann, Andrew R. Hudson, Soham Sinha, Jianyi Du, Daniel J. Shiwarski, Anastasiia Masaltseva, Fredrik Samdal Solberg, Jonathan Pham, Jason M. Szafron, Sean M. Wu, Adam W. Feinberg, Mark A. Skylar-Scott, and Alison L. Marsden. Rapid model-guided design of organ-scale synthetic vasculature for biomanufacturing. *Science*, 388(6752):1198–1204, 2025. doi: 10.1126/science. adj6152. URL https://www.science.org/doi/abs/10.1126/science.adj6152.

Arlo Sheridan, Tri M. Nguyen, Diptodip Deb, Wei-Chung Allen Lee, Stephan Saalfeld, Srinivas C. Turaga, Uri Manor, and Jan Funke. Local shape descriptors for neuron segmentation. *Nature Methods*, 20(2):295–303, Feb 2023. ISSN 1548-7105. doi: 10.1038/ s41592-022-01711-z. URL https://doi.org/10.1038/s41592-022-01711-z.

Suprosanna Shit, Johannes C. Paetzold, Anjany Sekuboyina, Ivan Ezhov, Alexander Unger, Andrey Zhylka, Josien P. W. Pluim, Ulrich Bauer, and Bjoern H. Menze. cldice - a novel topology-preserving loss function for tubular structure segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16560–16569, June 2021.

William Silversmith, J. Alexander Bae, Peter H. Li, and A.M. Wilson. Kimimaro: Skeletonize densely labeled 3D image segmentations, September 2021.

Carsen Stringer, Tim Wang, Michalis Michaelos, and Marius Pachitariu. Cellpose: a generalist algorithm for cellular segmentation. *Nature Methods*, 18(1):100–106, Jan 2021. ISSN 1548-7105. doi: 10.1038/s41592-020-01018-x. URL https://doi.org/10.1038/ s41592-020-01018-x.

Giles Tetteh, Velizar Efremov, Nils D. Forkert, Matthias Schneider, Jan Kirschke, Bruno Weber, Claus Zimmer, Marie Piraud, and Bjoern H. Menze. Deepvesselnet: Vessel segmentation, centerline prediction, and bifurcation detection in 3-d angiographic volumes, 2019. URL https://arxiv.org/abs/1803.09340.

Mihail Ivilinov Todorov, Johannes Christian Paetzold, Oliver Schoppe, Giles Tetteh, Suprosanna Shit, Velizar Efremov, Katalin Todorov-Völgyi, Marco Düring, Martin Dichgans, Marie Piraud, Bjoern Menze, and Ali Ertürk. Machine learning analysis of whole mouse brain vasculature. *Nature Methods*, 17(4):442–449, Apr 2020. ISSN 1548-7105. doi: 10.1038/s41592-020-0792-1. URL https://doi.org/10.1038/s41592-020-0792-1.

Engin Türetken, Christian Blum, Germán González, and Pascal Fua. Reconstructing geometrically consistent tree structures from noisy images. In Tianzi Jiang, Nassir Navab, Josien P. W. Pluim, and Max A. Viergever, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010*, pages 291–299, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg. ISBN 978-3-642-15705-9.

Engin Türetken, Germán González, Christian Blum, and Pascal Fua. Automated reconstruction of dendritic and axonal trees by global optimization with geometric priors. *Neuroinformatics*, 9(2):279–302, Sep 2011. ISSN 1559-0089. doi: 10.1007/s12021-011-9122-1. URL https://doi.org/10.1007/s12021-011-9122-1.

Engin Türetken, Fethallah Benmansour, Bjoern Andres, Przemysław Głowacki, Hanspeter Pfister, and Pascal Fua. Reconstructing curvilinear networks using path classifiers and integer programming. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(12):2515–2530, 2016. doi: 10.1109/TPAMI.2016.2519025.

Konrad W. Walek, Sabina Stefan, Jang-Hoon Lee, Pooja Puttigampala, Anna H. Kim, Seong Wook Park, Paul J. Marchand, Frederic Lesage, Tao Liu, Yu-Wen Alvin Huang, David A. Boas, Christopher Moore, and Jonghwan Lee. Near-lifespan longitudinal tracking of brain microvascular morphology, topology, and flow in male mice. *Nature Communications*, 14(1):2982, May 2023. ISSN 2041-1723. doi: 10.1038/s41467-023-38609-z. URL https://doi.org/10.1038/s41467-023-38609-z.

Bastian Wittmann, Yannick Wattenberg, Tamaz Amiranashvili, Suprosanna Shit, and Bjoern Menze. vesselfm: A foundation model for universal 3d blood vessel segmentation. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pages 20874–20884, June 2025.

Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable {detr}: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=gZ9hCDWe6ke.
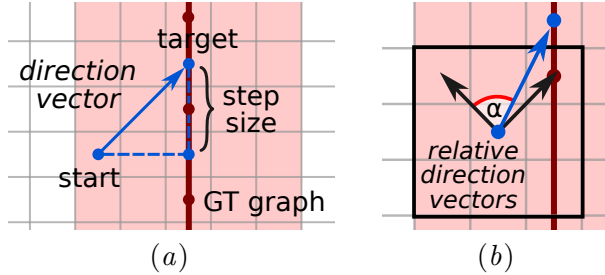
(a) (b)

Figure 4: **Direction vectors generation and angular difference penalty used in modified TEASAR.** (a) Direction vectors (blue) are generated by first identifying the closest point on the ground truth graph (yellow) and then stepping a fixed distance toward the root along the graph. Since the step size is constant across all voxel locations, direction vectors near the centerline exhibit smaller magnitudes, while those near the vessel boundary have larger magnitudes. (b) A penalty is assigned based on the angular difference between the predicted direction vector at the current voxel and the relative direction vector to each of its neighboring voxels (i.e., the "walk direction"). Lower penalties correspond to stronger alignment between the predicted and actual tracing directions, encouraging paths that follow the learned vessel orientation.

Appendix A. Extended Methods

A.1. Automated Root Detection

Vesselpose extracts a centerline graph using an adaptation of the TEASAR algorithm, which requires an initial set of vessel roots. In practice, these roots are annotated manually, a procedure that is both slow and labor-intensive. An automated strategy would remove this bottleneck, but detecting roots directly from the binary foreground mask is unreliable.

The voxel-wise direction field predicted by Vesselpose provides a direct workaround: roots correspond to sinks of this 3D flow field. A virtual particle is hereby initialized at every foreground voxel p whose distance to the background is at least $r_{\min}$. The particle position is set to $x_0 := p$ and updated iteratively by $x_i := x_{i-1} + \lambda \, v_{i-1}$, where v_{i-1} is the interpolated direction vector at x_{i-1} and $\lambda > 0$ is a step-size parameter. A position x_i is considered a sink when the displacement magnitude $\lambda \|v_i\|$ falls below a tolerance $\tau > 0$. After N iterations, all detected sinks are collected and used as candidate vessel roots for initializing the adapted TEASAR procedure.

To evaluate the effectiveness of automated root detection we apply this method to the three crops of the validation set of the Micro-CT Heart Data and report how many manually annotated ground truth roots are correctly detected. A root is considered to be correctly detected if it lies within a distance of 2 pixels from a calculated sink. For our experiments we further chose the following parameters: number of steps $N = 50$, step size $\lambda = 1.0$, tolerance $\tau = 0.1$ and min-radius threshold $r_{\min} = 3.0$.

As a result, the automated root detection algorithm is able to detect *all* of the ground truth roots present in all three volume samples, while producing additional false positive

18

Algorithm 1 Direction Vector Generation

Input : Vessel segmentation mask M and the corresponding skeleton graph $G = (V, E)$ with each edge $e \in E$ having a radius $r_e > 0$ assigned

Parameters: `step_size`> 0

Output: `vector_field` of direction vectors

Fit G to M by rescaling

Find leaf nodes in G

Select the root of each connected component to be the leaf node with maximal radius

Direct the edges in each connected component tree such that they are directed towards the root.

foreach `foreground_voxel` *of* M **do**
   Find `nearest_edge` $\in E$ to `foreground_voxel` in G
   Find `closest_point` to `foreground_voxel` on `nearest_edge`
   $d \leftarrow$ distance between `foreground_voxel` and `closest_point`
   if $d \leq r_{nearest\_edge}$ **then**
      `target_point` $\leftarrow$ move `closest_point` for `step_size` along G towards the root
      `direction_vector` $\leftarrow$ `target_point` - `foreground_voxel`
      Store `direction_vector` in `vector_field`
   end
end

return `vector_field`

---

roots. However, the additional false positive roots do not pose a problem to the TEASAR-based centerline generation since these can be filtered out during multi-root processing as described in 2.2.

A.2. Adaptive masking

In standard TEASAR to exclude the processed area a simple linear thresholding with fixed scale and const. However, in vesselpose these parameters vary with respect to the local radius as described in 2.2. Given predefined parameter ranges for $scale \in [s_{\min}, s_{\max}]$ and $const \in [c_{\min}, c_{\max}]$, and radius bounds $[r_{\min}, r_{\max}]$, we compute the normalized radius fraction

$$\alpha_r = \frac{r - r_{\min}}{r_{\max} - r_{\min}} \tag{4}$$

The scale and constant are then interpolated:

$$\text{scale}(r) = s_{\min} + \alpha_r(s_{\max} - s_{\min}) \ , \ \text{const}(r) = c_{\min} + \alpha_r(c_{\max} - c_{\min}) \tag{5}$$

The adaptive masking distance is finally defined as

$$d(r) = \text{scale}(r) \cdot r + \text{const}(r) \tag{6}$$

A.3. False split postprocessing

Finally, components without an assigned root are evaluated for potential merging with nearby trees. For each node in a disconnected component (the current node), we first
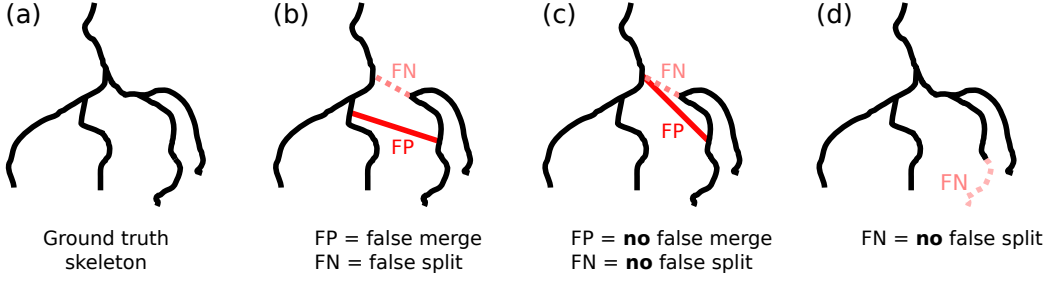
Figure 5: **False Merges & False Splits.** (a) A ground truth skeleton next to three possible predictions. (b) The predicted graph has one FN and one FP edge. The FP is a false merge since it connects two nodes which are not ancestor of each other. Consequently, the FN is a false split. (c) The FP edge is *not* a false merge since it keeps the ancestor relation w.r.t. to its parent node intact. Consequently, the FN is *not* a false split. (b) The predicted graph only has FN edges which are *not* false splits since they do not change the node ancestor relations.

identify neighboring nodes within a spatial distance of 5 voxels. Among these candidates, we retain only those that belong to a different tree and then compute their radius difference and angular deviation (based on the direction vectors) with respect to the current node. If the radius difference is below 3 and the angular difference is below 100 degrees, and adding an edge between the two nodes does not introduce a cycle, we connect them. In cases with multiple valid neighbors, we select the closest one. This step helps to reduce false splits introduced by the segmentation. However, very small or isolated components that do not meet these criteria may remain disconnected.

Appendix B. Extended Evaluation

B.1. Hierarchical Matching

In Algorithm 2 we demonstrate the steps involved in our hierarchical matching which takes the node label and its semantic into consideration rather than just the spatial proximity.

B.2. Formal definition of edge-wise F1-score for graph matching

Given the ground truth and prediction graphs G and P, along with a one-to-at-most-one node matching $\Phi : V_G \rightarrow V_P \cup \{\emptyset\}$, we evaluate their similarity by counting true positive (TP), false negative (FN), and false positive (FP) edges as follows.

- An edge (v, v') in G is a TP if and only if $(\Phi(v), \Phi(v'))$ is an edge in P.

- An edge (v, v') in G is a FN if and only if v and v' were matched and $(\Phi(v), \Phi(v'))$ is *not* an edge in P.

- An edge $(\Phi(v), \Phi(v'))$ in P is a FP if and only if (v, v') is *not* an edge in G.

Algorithm 2 Hierarchical matching between G and P

Input : Directed acyclic graphs G and P; Each node in each graph labelled as belonging to semantic class "root", "leaf", "branching point" or "intermediate point"; Each node in each graph has an associated 3d position; maximum distance for to points being matched d_{max}

Output: `match_dict`

Initialize empty `match_dict`

foreach *node v in G* **do**

    Determine the set W_v of closest nodes in P that lie within distance d_{max} to v

    Determine subset $W_{v,sem} \subseteq W_v$ of nodes with same semantic class as v

    Sort each subset by distance to v

    Append to form sorted list $W_v^{sorted} = [W_{v,sem},\ W_v \backslash W_{v,sem}]$

end

Sort root nodes r of G analogously, i.e.: Primary sorting criterion: non-empty $W_{r,sem}$ before all others; secondary criterion: distance of closest $w \in W_r$

foreach *r in sorted list of root nodes in G* **do**

    foreach *class label c in sorted list $["root", "branching point" or "leaf", "intermediate point"]$*

    do

        foreach *node v with class label c in depth-first traversal of G from r* **do**

            if *class label of v is "root"* **then**

                remove all elements w in W_v^{sorted} if parent of w is already matched to another tree than r

            end

            if *W_v^{sorted} is not empty* **then**

                if *class label of v is no "root"* **then**

                    get mask m_v with true elements if parent of $w \in W_v^{sorted}$ is matched to same tree as r

                    if *$W_v^{sorted}[m_v]$ is not empty* **then**

                        update W_v^{sorted} with $W_v^{sorted}[m_v]$

                    end

                end

                pick first w in W_v^{sorted}

                store (v, w) in `match_dict`

                foreach *not yet visited node v' in G* **do**

                    remove w from $W_{v'}^{sorted}$

                end

            end

        end

    end

end

return `match_dict`

The *edge-wise F1-score* is defined as

$$F_1^{\text{edge}} := \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

B.3. Related work: Commonly used metrics

A unified approach for comparing graph structures is still lacking Lyu et al. (2022), and numerous evaluation strategies have emerged across the biomedical literature. On one hand, there are *detection- and segmentation-based metrics* for comparing centerlines, branching points, or graph edges, which we find unsuitable for graph-based evaluation: For example, clDice (Shit et al., 2021) does not reflect topology, as it operates at the pixel level and fails to penalize structural changes like loops or disconnections—errors that may drastically alter topology but only minimally impact clDice scores as few pixels are missing or added. The same limitation applies to precision, recall, and F1 scores when applied on pixel level. If mean average precision (mAP)(Lin et al., 2014) is used, as in Prabhakar et al. (2024); Naeem et al. (2024), nodes and edges are compared based on their overlap of the bounding boxes. However, as shown in Foucart et al. (2023), intersection-over-union (IoU) is not well-suited for small objects, especially in 3d, where meeting high IoU thresholds becomes impractical.

On the other hand, *graph similarity measures* are commonly used in related studies (Prabhakar et al., 2024; Naeem et al., 2024; Drees et al., 2019, 2021). However, the Street Mover's Distance (SMD) does not preserve connectivity information as it converts graph to point clouds for distance computation. Additionally, SMD is sensitive to resampling and hyperparameters, making it less reliable for topological evaluation. In contrast, Betti numbers provide a topological perspective, but, e.g., Betti-0 is too coarse, as it does not capture topological errors within individual trees (Lux et al., 2025). Instead, precision, recall, and F1 scores—when applied at the graph level, particularly for edges, as in Drees et al. (2019, 2021)—provide meaningful insights into topological correctness. However, these metrics do not account for the structural impact of individual errors: missing or spurious edges may vary greatly in severity depending on how drastically they alter the topology of the graph, a nuance not captured by the F1 score.

Another family of metrics includes *tree edit distance (TED)-based measures* like Li et al. (2023); Matula et al. (2015), which quantify the number of operations (e.g., node or edge insertions/deletions) needed to transform the predicted graph into the ground truth. These metrics offer an intuitive and meaningful notion of similarity, but we have not seen them commonly used in related work. For example, Li et al. (2023) introduces a TED variant specifically for vasculature graphs, but it penalizes heavily false merges and splits between two trees as it penalizes the falsely merged tree twice by adding the counts for removing every falsely added node and edge; and then for creating the missing tree from scratch. Another intuitive measure is the TRA metric (Matula et al., 2015), originally designed for evaluating cell lineages, but it is not directly applicable here as it relies on IoU-based matching.

Our observations are consistent with Lyu et al. (2022), who conclude that selecting a reliable and unbiased evaluation metric remains an open problem in the community. With this work, we aim to contribute to this discussion by proposing a greedy hierarchical matching strategy and introducing false splits and false merges as topology-aware measures for robust assessment of vessel graph reconstructions. Nevertheless, we believe that a more systematic analysis of how node matching and sampling choices influence evaluation outcomes

Table 3: Quantitative comparison of our hierarchical matching with greedy one-to-one matching and optimal Hungarian matching. Lower false-merge (FM) and false-split (FS) values indicate better topology preservation during matching. Results are shown for Vesselpose on the validation set of the Multi-Tree Synthetic dataset. For this analysis, graphs were resampled to include only roots, branching points, and end nodes.

Matching	Edges			FM		FS	
	F1↑	Prec↑	Rec↑	Rel.↓	Abs.↓	Rel.↓	Abs.↓
Greedy	0.8	0.81	0.79	0.01	49	0.01	48.6
Hungarian	**0.81**	**0.82**	**0.80**	0.02	62.6	0.01	62.3
Hierarchical (Ours)	0.80	0.81	0.79	**0.007**	**29.7**	**0.007**	**29.3**

is still needed to establish common recommendations—an important direction for future work, but beyond the scope of this paper.

Appendix C. Extended Experiments

C.1. Single-Tree Synthetic Data

The dataset comprises 500 volumes of size 256^3 voxels and is split into training, validation, and test sets. The training, validation and test split contain 368, 32 and 100 samples respectively. The network is trained on randomly sampled input patches of size 128^3 voxels with intensity shift augmentation.

C.2. Parse 2022 Challenge

The data has a in-plane size of 512×512 pixels and its z-stack comprises between 295 and 390 slices. The training, validation and test split contain 72, 8 and 20 samples respectively. Following Naeem et al. (2025), all volumes are resampled to an isotropic resolution of $0.5mm$. For both training and inference, we use input sizes of size 256^3 voxels and do not apply any data augmentation.

C.3. Multi-tree Synthetic Data

This dataset is obtained from (Tetteh et al., 2019). Each volume measures $325 \times 304 \times 600$ voxels. We follow the information provided in Naeem et al. (2024); Prabhakar et al. (2024) about how many samples where used in which data split and utilized the first 50 volumes. The training, validation and test split contain 37, 3 and 10 samples respectively. Training and inference are performed with a input size of 256^3 voxels, using intensity shifts and masked out crops as data augmentations.
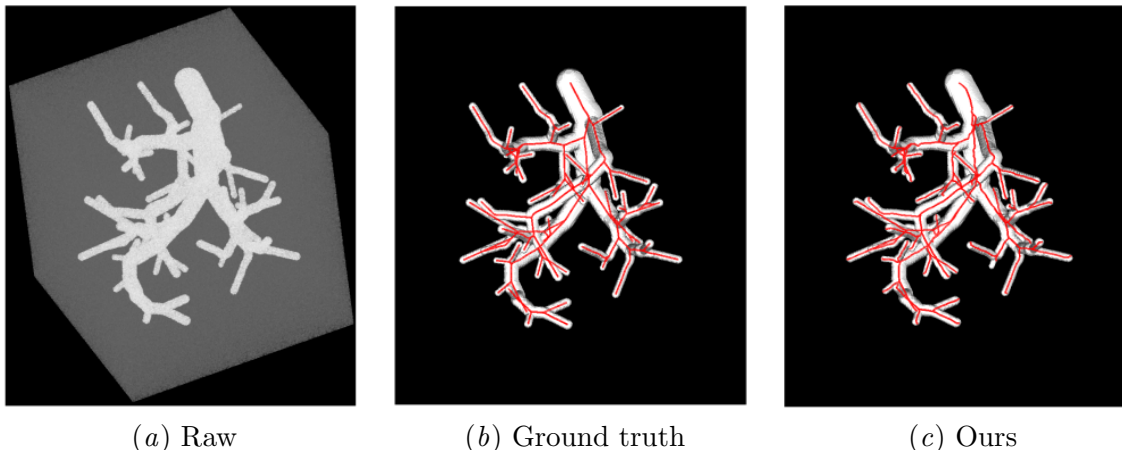
(*a*) Raw   (*b*) Ground truth   (*c*) Ours

Figure 6: **Qualitative comparison for single-tree synthetic.** A 3D rendering of one of the samples. (a) shows the raw image (b) Ground-truth skeletons overlaid on the segmentation mask (c) our predicted skeleton overlaid on the predicted binary segmentation. Our method (c) produces a reconstruction that closely matches the ground truth(b), capturing fine structures and maintaining topological consistency.

C.4. Micro-CT Heart data

This dataset contains samples from both preeclamptic and healthy animals. The individual samples measur approximately $1300 \times 1300 \times 1700$ voxels at a resolution of $12\mu m$. To obtain foreground masks for fine-tuning the U-Net on the micro-CT heart data, we train a random forest classifier (Breiman, 2001) with 100 trees and a maximum depth of 10. We include Frangi vesselness features (Frangi et al., 2000) as additional input features for the random forest classifier and manually annotate a subset of vessels as ground truth. We then used the resulting foreground masks to fine-tune the U-Net model which was pre-trained with the synthetic multi-tree dataset. We fine-tune the U-Net by freezing all but the final layer to predict the foreground mask.

A key challenge associated with this dataset is its large size and the presence of substantial heart chambers, which occupy much of the volume. Thus, we downsample the data by a factor of 0.5 in each dimension and apply a heuristic approach to remove the chambers by eliminating the largest connected component in each 2D segmentation slice. Fine-tuning and inference are performed with a input size of 256^3 voxels.

C.5. Ablation experiments

To assess the contribution of each modification to the standard TEASAR algorithm, we conducted a systematic ablation study. Starting from the kimimaro TEASAR, we incrementally added each proposed component in separate experiments. We observe a consistent improvement in performance with every addition in Table 4, ultimately achieving the lowest FM and FS values with our full model configuration.

(a) Raw   (b) Ground truth   (c) Ours
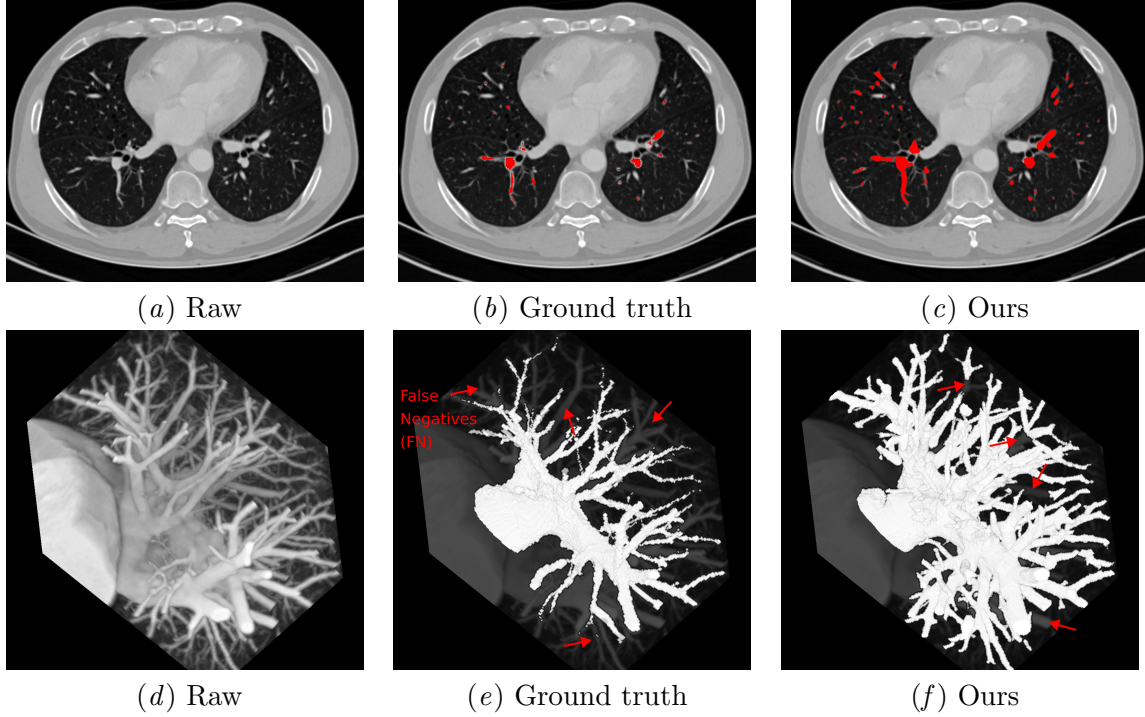
(d) Raw   (e) Ground truth   (f) Ours

Figure 7: **Ground truth segmentation misses small vessels.** First row: A 2D slice from one sample. (a) shows the raw image, while (b) and (c) show the raw image overlaid with the provided ground truth segmentation and our segmentation, respectively. Second row: A 3D crop from the same sample. (d) shows the raw image, and (e) and (f) show the raw image overlaid with the ground truth and our segmentation, respectively. Both segmentation masks miss vessel segments visible in the raw data (highlighted with red arrows in (e) and (f)) and contain disconnected components. However, our method recovers more of the fine, smaller vessels. These observations indicate that the provided ground truth segmentation and skeletonizations derived from it are not well suited as a benchmarking ground truth.

(a) Ground truth $\qquad$ (b) Ours $\qquad$ (c) U-Net + TEASAR

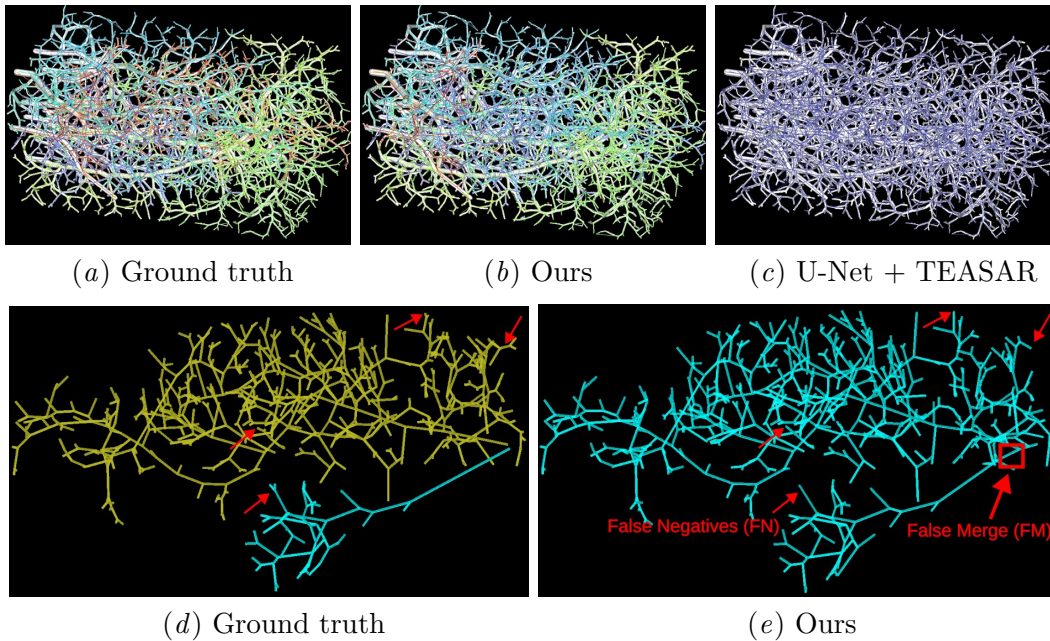(d) Ground truth $\qquad\qquad\qquad$ (e) Ours

Figure 8: Qualitative results for the multi-tree synthetic dataset. First row: Segmentation mask and skeletons overlaid, where each color represents a distinct tree. Our approach separates most trees, whereas U-Net and TEASAR merge all trees into one component. Second row: Failure cases for our method, including missed small terminal branches (red arrows) and falsely merged trees (red rectangle).

(a) Ground Truth $\qquad\qquad$ (b) Vesselpose $\qquad\qquad$ (c) Baseline
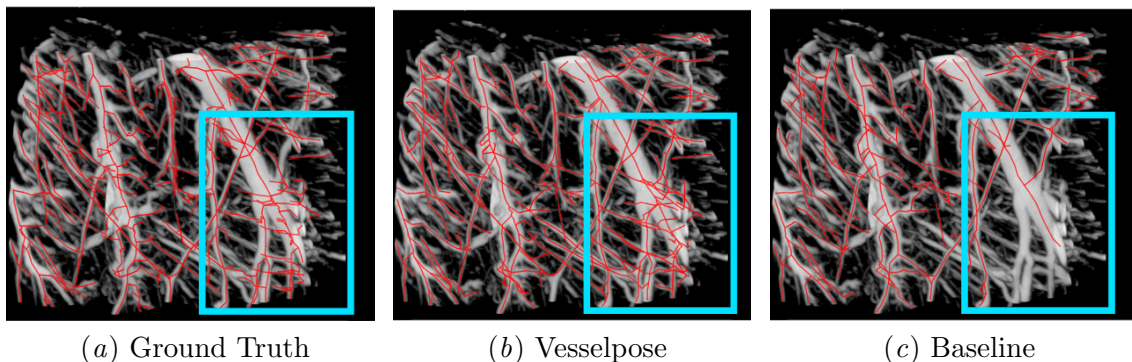
Figure 9: **Qualitative results of the micro-CT data.** Illustrated is one 3D annotated crop from the raw micro-CT test data together with varying vessel skeleton graphs in red: (a) shows the annotated ground truth skeleton; (b) shows the results of our proposed method; (c) shows the result of the baseline, which consists of a U-Net for foreground segmentation followed by the original TEASAR algorithm. Overall, our method accurately captures the vessel structures and aligns well with the ground truth. In contrast, the baseline method fails to trace many vessel branches, in particular in the highlighted region within the blue box.

26

Table 4: Ablation study illustrating the contribution of individual components of our method and how they incrementally improve performance over a U-Net with standard TEASAR (Ronneberger et al., 2015; Sato et al., 2000). We add the following components step by step: support for multiple roots per connected component (multi-root); an additional penalty for tracing along vectors with small magnitudes (vec mag); an additional penalty for tracing in the same direction as the direction vector (vec dir); and adaptive masking to mark processed regions (adapt. mask). Results are shown for Vesselpose on the validation set of the Multi-Tree Synthetic dataset. For this analysis, graphs were resampled to include only roots, branching points, and end nodes.

Experiments	multi root	vec mag	vec dir	adapt. mask	Edges			FM	FS
					F1↑	Prec↑	Rec↑	Abs.↓	Abs.↓
UNet+TEASAR	✗	✗	✗	✗	0.77	0.90	0.68	48.3	41.3
Ours	✓	✗	✗	✗	0.83	0.91	0.75	38.3	38.0
	✓	✓	✗	✗	0.82	0.91	0.75	37.3	37.6
	✓	✓	✓	✗	0.85	0.89	0.82	33.6	35.0
	✓	✓	✓	✓	0.88	0.89	0.87	29.6	30.6