Supplementary Material

Cross-domain and Cross-dimension Learning for Image-to-Graph Transformers

6. Additional ablation studies

Additionally to our main experiments, we present ablation studies on further aspects of our proposed framework. These ablation studies give a deeper understanding of the components' dynamics and guide future reimplementations and adaptions.

6.1. Generalizability

In Table 4, we present the results of our experiment E) (see Table 1) with one additional configuration. In this configuration, we pretrain the model on the synthetic OCTA dataset [35] instead of the U.S. cities dataset [22]. This pretraining strategy outperforms all baselines and increases the performance compared to our main experimental setting (see Section 4). We attribute this improvement to the smaller domain gap between the new source domain (i.e., retinal blood vessels) and the target domain (i.e., a mouse's cerebrovasculature). These results show how our method generalizes seamlessly to new domains. Furthermore, they substantiate the rationale behind our experimental design: by showcasing the utility of our method in a challenging setting, focused on the most intricate transfer learning scenarios, we establish its effectiveness in more straightforward transfer learning situations (as presented in Table 4 as well.

6.2. Regularized edge sampling loss

Next, we conduct an ablation study on the effect of the regularized edge sampling loss. As explained in Section 4, the new loss stabilizes training and increases convergence speed. This effect is shown in Figure 4, where the training loss decreases faster from the beginning on and convergences towards a lower level compared to the baseline loss formulation. This effect can be observed across all datasets and training strategies. Also, we experiment with different foreground-to-background-edge ratios r (see Section 3.1). Table 5 shows that the performance stays stable across a large range of r-values. These results underline our hypothesis from Section 3.1 that $\mathcal{L}_{Resln}$ reduces the hyperparameter space because it does not require careful optimization.

Furthermore, we study the effect of different edge-sampling strategies on our loss formulation in Table 6. Specifically, we compare our fixed-ratio upsampling strategy with a varying-r upsampling (i.e., for each batch, we randomly choose r with a uniform distribution in $(,1]$), and a fixed-ratio subsampling strategy. The decreased performance with a varying-r upsampling strategy shows that a fixed r is important for our loss formulation. We further find that subsampling is a valid alternative in scenarios where data is extremely scarce (e.g., Experiment A) but performs worse when more data is available (e.g., Experiment E). Notably, Shit et al. [42] proposed a one-sided subsampling strategy, i.e., subsampling only the background edges if the ratio is **above** a certain ratio. This strategy is problematic when the target dataset contains dense graphs, in which our loss formulation upsamples the background edges (see Table 7 for dataset statistics). Furthermore, the official relationformer repository does use a dynamic subsampling strategy but selects background edges up to an absolute threshold m, which introduces strong hyperparameter sensitivity. Table 7 shows that up- or sub-sampling only one edge type (e.g., the background edges) would not be sufficient.

6.3. Domain adaptation framework

Table 8 shows an ablation study of our domain adaptation framework's components. $\mathcal{L}_{img}$, $\mathcal{L}_{graph}$, and $\mathcal{L}_{cst}$ refer to the optimization terms from Section 3.2. Using the image-level alignment alone already yields a performance increase of around 30 % compared to not using our framework at all. We attribute this observation to the large image-level differences between the source and target domain, which hinders knowledge transfer in the feature extractor if an adversarial does not mitigate it. The graph-level adversarial slightly decreases the performance when being applied without consistency regularization (i.e., $\mathcal{L}_{cst}$). This decrease is likely caused by the abstraction level of the transformer's tokenized graph representation. Without any further guidance (e.g., by the image-level domain classifier through consistency regularization), the graph-level classifier does not provide a precise gradient toward a domain-invariant representation. Combining all three components yields the best results, supporting our hypothesis that the graph-level adversarial needs regularization by the image-level adversarial.

Furthermore, we study the impact of our projection function and loss formulation without applying our domain adaptation framework. Table 9 shows that our other contributions alone enable transfer learning across dimensions. This enables transfer learning without access to the target domain during pretraining. However, even in these cases, our DA framework yields the best performance. Table 2 shows a similar trend in cases without dimension shift.

6.4. Adversarial learning coefficient

In Table 10, we ablate on the domain adversarial learning coefficient α. α is the factor with which the gradient in the GRL is multiplied before passing it to the respective model component, i.e., the feature extractor for the image-level adversarial and the encoder-decoder for the graph-level adversarial (see Section 3.2). We use the α schedule proposed by Chen et al. [10], which increases α during the training until reaching a fixed maximum. Table 10 shows that the right choice of α is crucial and that a suboptimal value can decrease downstream performance. We attribute this observation to the model's tradeoff between learning to produce domain-invariant features (i.e., domain confusion) and task learning (i.e., graph extraction). If α is too large, the adversarial loss dominates the task loss, and the network does not learn how to produce meaningful features. If it is too small, the domain gap between the source and target domain stays too large, and knowledge transfer is impeded. Figure 5 shows how a small α (e.g., $\alpha = 0.3$) is not sufficient to learn domain-invariant features while an α-value that is too large does not increase domain confusion but obstructs learning the core task. Note that the specific α value must be optimized for the used datasets and is not domain-invariant.

6.5. Target dataset size

Lastly, Figure 6 shows the results of an ablation study on the target dataset size. We plot the harmonic mean of node and edge

Table 4. Ablation study on the pretraining dataset. As the domain gap between the source and target domain decreases, the downstream performance increases. We show that our method is generalizable across different pretraining datasets.

Fine Tuning Training Set	**(Pre-)Training Strategy**	**Node-mAP↑**	**Node-mAR↑**	**Edge-mAP↑**	**Edge-mAR↑**	**SMD ↓**
Microscopic images [47]	No Pretraining [18]	0.231	0.308	0.249	0.329	0.017
	Self-supervised [9]	0.344	0.404	0.363	0.425	0.017
	Supervised	⋆	⋆	⋆	⋆	⋆
	Ours, pretr. on cities	0.483	0.535	0.523	0.566	0.017
	Ours, pretr. on OCTA	**0.548**	**0.583**	**0.588**	**0.615**	**0.016**

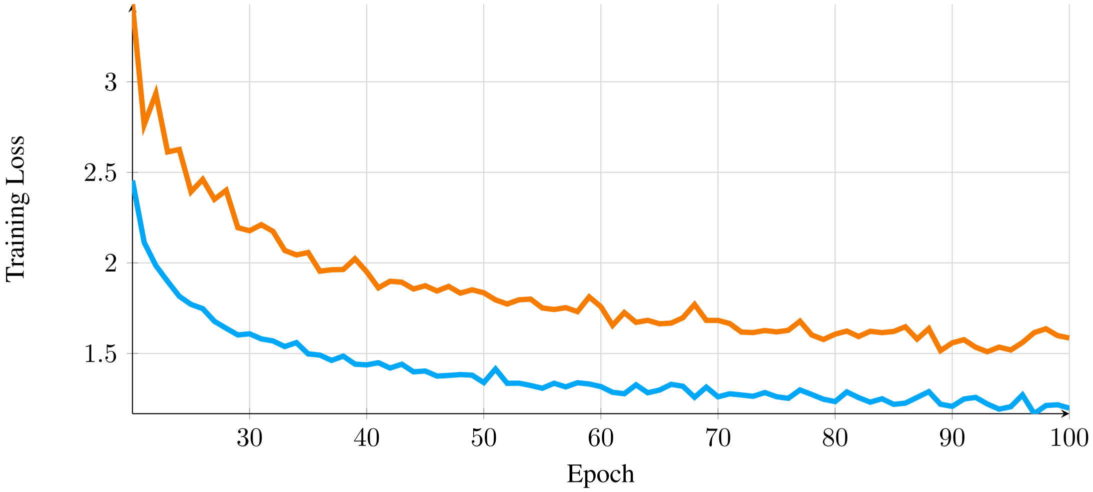

Figure 4. Training loss curves. The orange line depicts the training loss without our regularized edge sampling loss $\mathcal{L}_{Resln}$ and the blue line with $\mathcal{L}_{Resln}$, respectively. $\mathcal{L}_{Resln}$ shows faster convergence from the beginning on.

Table 5. Ablation study on the loss ratio r for $\mathcal{L}_{Resln}$ as described in Section 3.1 (experiments congruent to Table 1). We observe that $\mathcal{L}_{Resln}$ is stable across varying loss ratios and does not require sensitive hyperparameter tuning.

Experiment	r	**Node mAP**	**Node mAR**	**Edge mAP**	**Edge mAR**
D	0.05	0.3539	0.4446	0.2166	0.3102
	0.1	**0.3564**	**0.4498**	0.2209	0.3218
	0.15	0.3470	0.4380	0.2164	0.3122
	0.2	0.3532	0.4449	0.2203	0.3193
	0.3	0.3451	0.4351	0.2183	0.3153
	0.5	0.3470	0.4407	0.2231	0.3242
	0.8	0.3462	0.4394	**0.2253**	**0.3288**

mAP (see Section 4) of our method and the *no-pretraining* baseline against the size of the target dataset. We observe that our method consistently outperforms the baseline across all dataset sizes. However, as the number of samples increases, the performance difference between the two methods decreases. This observation is expected because transfer learning becomes less effective (and is also less required) when enough target domain samples are available. Our framework is especially useful if target data is scarce.

7. Model & training details

To find the optimal hyperparameters, we follow a three-step approach. First, we optimize the model architecture hyperparameters (e.g., model size) with a random weight initialization (i.e., no transfer learning) on the target task. Then, we fix these hyperparameters for the remainder of the optimization process. An overview of the model hyperparameters for each experiment can be found in Table 11. Second, we optimize the training hyperparameters (e.g., learning rate or batch size) for pretraining on the source task with the fixed model architecture hyperparameters from step 1. Third, we use the pretrained model with the best performance on the source task and optimize the training parameters on the target task for each training strategy separately on the validation set. We follow this approach because optimizing the whole pipeline (including pretraining and fine-tuning) in a brute-force manner would require too many resources in terms of computational power and energy consumption. Table 12 depicts the training hyperparameters for the target task for all the experiments listed in Section 4.

Table 6. Ablation study on different edge sampling strategies.

Strategy	**Experiment A**		**Experiment E**	
	node-mAP	**edge-mAP**	**node-mAP**	**edge-mAP**
subsampling	0.172	0.125	0.237	0.277
varying-r	0.156	0.115	0.218	0.134
oversampling (ours)	0.173	0.129	0.267	0.323

Table 7. Edge statistics for the used datasets. The varying ratios between active and background edges underline the utility of our dynamic loss formulation. Different datasets require upsampling for active and background edges.

Dataset	**Avg. Edges**	**Avg. edge ratio**	**Upsampling background**	**Upsampling active**
20 U.S. Cities	6.37	0.53	92.5%	7.5%
Agadez	5.05	0.75	96.2%	3.8%
Munich	4.69	0.77	97.1%	2.9%
Synth. OCTA	32.57	0.05	1.4%	98.6%
OCTA-500	11.07	0.18	47.5%	52.5%
Synth. MRI	22.37	0.07	0.3%	99.7%
Microscopy	33.31	0.05	0.3%	99.7%

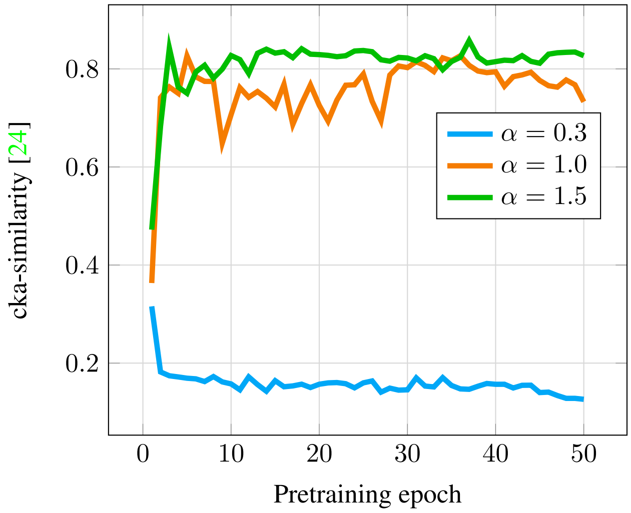

Figure 5. cka-similarity [24] (y-axis) between the feature representations of source and target domain during pretraining. $alpha$ must be sufficiently large such that the similarity increases during training. From a certain threshold on, the similarity does not increase further. We associate a high similarity between both domains with the model learning domain-invariant features.

8. Datasets

In the following, we describe the properties and sampling of our six diverse image datasets and the unlabeled datasets we used for the self-supervised baseline.

8.1. Training set - 20 U.S. Cities

[22] is a city-scale dataset consisting of satellite remote sensing (SRS) images from 20 U.S. cities and their road graphs covering a total area of 720 km^2. The satellite images are retrieved in the RGB format via the Google Maps API [19]. The corresponding road network graphs are extracted from OpenStreetMap [20]. We cut the resulting images and labels into overlapping patches of 128x128 pixels with a spatial resolution of one meter per pixel. In these patches, we eliminate redundant nodes (i.e., nodes of degree 2 with a curvature of fewer than 160 degrees) to simplify the prediction task [4].

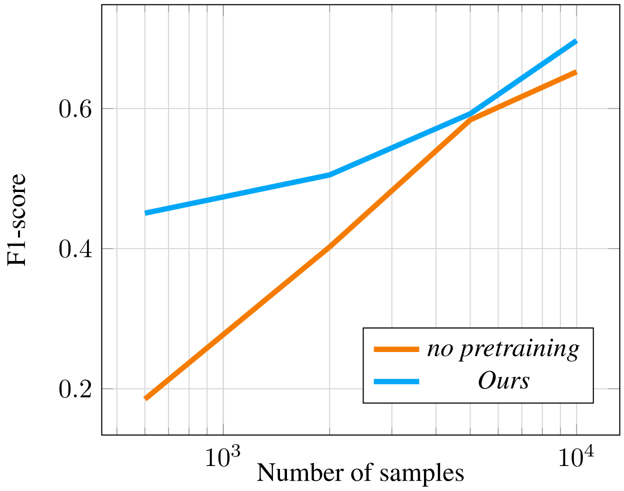

Figure 6. F1-scores (y-axis) over different target dataset sizes (x-axis). The F1-score is calculated between the node and edge mAP as described in Section 9. The orange line depicts the F1-scores of the *no-pretraining* baseline, and the blue line with our contributions, as described in Section 4. The x-axis is in logarithmic scale. We observe that our contributions are significantly reducing data requirements, especially when data is scarce.

8.2. Agadez and Munich, cities around the globe

We create our own image dataset from OpenStreetMap[2] covering areas that differ from those covered by the 20 U.S. cities dataset in terms of geographical and structural characteristics. Geographical characteristics refer to the area's natural features (e.g., vegetation), while structural characteristics relate to anthropogenic (human-made) structures that affect an area's surface (e.g., street type) or layout (e.g., city type). The complete dataset contains a 4 km^2 area of 11 cities with different characteristics in different parts of the world. Both source images and labels were obtained in the same manner as for the 20 U.S. cities dataset [22]. Our dataset is

[2] https://www.openstreetmap.org

Table 8. Ablation study on our domain adaptation framework in a transfer learning setting (experiments congruent to Table 1). $\mathcal{L}_{img}$, $\mathcal{L}_{graph}$, and $\mathcal{L}_{cst}$ refer to the optimization terms from Section 3.2. We find that performance improvements are associated with all adaptation components. Using the complete optimization term as presented in Section 3.3 yields the best results.

Experiment	$\mathcal{L}_{img}$	$\mathcal{L}_{graph}$	$\mathcal{L}_{cst}$	Node mAP	Node mAR	Edge mAP	Edge mAR
C	✗	✗	✗	0.3423	0.4341	0.2581	0.3414
	✓	✗	✗	0.4286	0.5073	0.3293	0.4264
	✗	✓	✗	0.3264	0.4136	0.2355	0.3117
	✓	✓	✗	0.4071	0.4980	0.2685	0.3831
	✓	✓	✓	**0.4909**	**0.5712**	**0.3656**	**0.4887**

Table 9. The performance of our contributions with and without our DA framework in the additional experiment from Table 4. We show that smaller domain gaps (here, from OCTA to microscopy images) can be bridged without our DA framework even with dimension shift.

Method	node-mAP	node-mAR	edge-mAP	edge-mAR
No Pretraining	0.231	0.308	0.249	0.329
No DA	0.508	0.549	0.551	0.584
Ours	0.548	0.583	0.588	0.615

Table 10. Ablation study on the domain adversarial learning coefficient α (experiments congruent to Table 1). α must be optimized such that the adversarial loss balances with the graph extraction loss. An α value of 0 is equivalent to not using the domain adaptation framework.

Experiment	α	Node mAP	Node mAR	Edge mAP	Edge mAR
C	0.0	0.3884	0.4755	0.2947	0.3767
	0.1	0.3123	0.4076	0.2258	0.3044
	0.3	0.3381	0.4287	0.2439	0.3240
	0.5	0.4563	0.5395	0.3614	0.4618
	0.8	0.4623	0.5458	0.3464	0.4687
	1.0	**0.4909**	**0.5712**	**0.3656**	**0.4887**
	1.5	0.3914	0.4726	0.2854	0.3897
	2.0	0.0530	0.1719	0.0214	0.0451

accessible in our GitHub repository [3].

For our experiments, we choose two cities, Agadez and Munich, whose characteristics differ from the 20 U.S. cities dataset in different aspects as displayed in Table 13. We strategically choose those cities to investigate how differences in specific characteristics between the source and target domain affect knowledge transfer and how transfer learning strategies should be adapted to these differences. We especially test our hypothesis that surface-level characteristics are captured by different components than layout-level characteristics. These new datasets enable the verification because Agadez differs from 20 U.S. cities in surface-level characteristics (e.g., vegetation, street type, and buildings) but shares a similar city layout (i.e., grid plan). Note that although Agadez has a historical city center, we chose a part of the city that follows the typical grid layout. Contrary to this, Munich is similar to U.S. cities in surface-level characteristics while following a different city layout (i.e., a historical European city layout). We test each city dataset separately.

8.3. Synthetic OCTA

The synthetic Optical Coherence Tomography Angiography (OCTA) dataset [35] consists of synthetic OCTA scans with intrinsically matching ground truth labels, namely the corresponding segmentation map and the vessel graphs. The images were created using a simulation based on the physiological principles of angiogenesis to replicate the intricate retinal vascular plexuses [40], followed by incorporating physics-based modifications to emulate the image acquisition process of OCTA along with the usual artifacts. We project the 3D OCTA images along the main axis, split them scan-wise between training and testing sets, and extract 2600 overlapping samples of 128 × 128 pixels. For training our self-supervised baseline, we use the same procedure on 200 additional synthetic OCTA scans to extract almost 100,000 patches.

8.4. OCTA-500

The OCTA-500 dataset [26] includes 300 OCTA scans with a 6 mm × 6 mm field of view. The 400 × 400 large *en-face* projection images were manually annotated with sparse vessel labels. We extract the graphs from these segmentation maps using the method presented by Drees et al. [14]. We split the scans patient-wise between training and testing sets and create around 3000 overlapping patches with a spatial size of 128 × 128. Furthermore, we combine the OCTA scans from *OCTA-500* with the scans of the *ROSE* dataset [32] to obtain around 40,000 patches for training our self-supervised baseline. This combination is necessary for an unlabeled dataset large enough for self-supervised pretraining.

8.5. Synthetic MRI

The Synthetic MRI dataset [46] is a synthetical 3D dataset that simulates the characteristics of clinical vessel datasets. The original dataset provides ground truth labels for vessel segmentation, centerlines, and bifurcation points. The ground truth graphs are obtained with the method described by Drees et al. [14]. We cut the volumes and their graphs in overlapping patches of $64 \times 64 \times 64$ voxels. We use the same dataset with all 80,000 patches for our self-supervised pretraining.

[3] GitHub repository will be made publicly available upon acceptance

Table 11. Model details for each experiment. The hyperparameters are the same for each model of the respective experiment. The latent space resolution is controlled via the CNN backbone's stride. It determines the feature size between the backbone and the transformer.

Target Set	Backbone		Latent Space	Transformer				FFN
	Type	**Hid. Dim.**	**Resolution**	**Hid. Dim.**	**# Lay.**	**# Obj Token**	**# RLN Token**	**Hid. Dim.**
Agadez	ResNet101	512	Multi-Level	512	3	80	2	1024
Munich	ResNet101	512	Multi-Level	512	3	80	2	1024
Synthetic OCTA	ResNet101	512	Multi-Level	512	3	80	5	1024
OCTA-500	ResNet101	512	Multi-Level	512	3	80	2	1024
Synthetic MRI	SeresNet	256	$2 \times 2 \times 2$	552	4	120	2	1280
Whole Brain Vessels	SeresNet	256	$2 \times 2 \times 2$	552	4	120	2	1280

Table 12. Training details and hyperparameters for each trained model.

Experiment	Batch Size	Epochs	Learning Rate		Loss Coeff.				
			Backbone	**Transformer**	λ_{gIoU}	λ_{cls}	λ_{reg}	λ_{Resln}	λ_{reg}
A	32	100	0.00002	0.0002	2	3	5	5	1.0
B	32	100	0.00002	0.0002	2	3	5	5	1.0
C	32	100	0.00002	0.0002	2	3	5	5	1.0
D	32	100	0.00007	0.00007	3	4	2	6	0.8
E	32	100	0.00007	0.00007	3	4	2	6	0.8

8.6. Whole Brain Vessels

The Whole Brain Vessel dataset [37] is a publicly available open graph benchmark dataset for link prediction (ogbl-vessel[4]). It consists of a graph representing the entirety of the mouse brain's vascular structure down to the capillary level. Todorov et al. [47] obtained the raw vessel scans using tissue-clearing methods and fluorescent microscopy and then segmented the brain vasculature using CNNs. The dataset has an image, segmentation, and graph representation. We create overlapping patches with a spatial size of $50 \times 50 \times 50$ voxels and remove artifactual patches (e.g., patches containing only noise). We extract 43,500 image patches from an unlabeled whole-brain mouse scan obtained with the vDISCO pipeline [16] for training our self-supervised model.

9. Evaluation metrics

We choose to evaluate our models' performance using three different evaluation metric types: 1) topological metrics, 2) graph distance metrics, and 3) object detection metrics.

Topological metrics The TOPO-score [6] samples multiple sub-graphs starting from different seed locations from the ground truth and measures its similarity to the inferred graph from the predicted graph with the same seed location. The similarity is measured by matching a fixed amount of points between the two graphs. Two points from two graphs are matched if the distance between their spatial coordinates is below a threshold. The result of this matching across all sampled subgraphs is used for calculating precision and recall. This method accurately quantifies a prediction's geometrical (i.e., the roads' geographical position) and topological (i.e., the roads' interconnections) quality. We use the implementation and parameters from Biagioni et al. [6]. These metrics are not implemented in 3D.

Graph distance metrics The street mover distance (SMD) approximates the Wasserstein distance between a fixed number of uniformly sampled points along the ground truth graph and the predicted graph. Intuitively, it represents the minimal distance by which the predicted graph must be moved to match the ground truth [4].

Object detection metrics Further, we resort to widely-used object detection metrics: mean average precision (mAP) and mean average recall (mAR) [36]. To calculate each detection's intersection over union (IoU), we create a hypothetical bounding box of fixed size around each node. Similarly, we create bounding boxes around the edges with a minimum spatial size m in all dimensions. This minimum holds for edges that connect two nodes a and b where the difference between the coordinates in one dimension is lower than m (e.g. if $|a_x - b_x| < m$). We calculate the mean AP and AR between the values of different IoU thresholds (i.e., 0.5 and 0.95).

10. Additional quantitative results

In Table 14, we present our main results from Table 1 in addition to the results' standard deviation across five mutually exclusive folds of the test set.

[4] https://ogb.stanford.edu/docs/linkprop/#ogbl-vessel

Table 13. Overview of the used datasets, selected characteristics, and respective training, validation, and test set sizes.

Dataset	Road Description				Vessel Description		Split		
	Street Type	Vegetation	Layout	Continent	Dimension	Spatial Size	Train	Val	Test
20 U.S. Cities [22]	Sealed	Rich	Grid-plan	N. America	2D	128×128	99.2k	24.8k	25k
Global Diverse Cities									
Agadez	Unsealed	Arid	Grid-plan	Africa	2D	128×128	480	120	290
Munich	Sealed	Rich	Historical	Europe	2D	128×128	440	110	220
Synth. OCTA [35]	-	-	-	-	2D	128×128	480	120	2k
OCTA-500 [26]	-	-	-	-	2D	128×128	1.6k	400	2.2k
Synth. MRI [40]	-	-	-	-	3D	64× 64 ×64	4k	1k	5k
Microscopy [47]	-	-	-	-	3D	50× 50 ×50	4k	1k	1.2k
Unlabeled datasets									
20 U.S. Cities	-	-	-	-	2D	128×128	124k	-	-
Synth. OCTA [35]	-	-	-	-	2D	128×128	96.4k	-	-
Real OCTA [26, 32]	-	-	-	-	2D	128×128	40k	-	-
Synthetic MRI [40]	-	-	-	-	3D	64× 64 ×64	80k	-	-
Microscopy [16]	-	-	-	-	3D	50× 50 ×50	43.5k	-	-

11. Additional qualitative results

We are providing additional qualitative results in the form of multiple figures; please see Figure 7 - 11.

Table 14. Main results with standard deviations. Quantitative Results for our cross-dimensional image-to-graph transfer learning framework. All models are pretrained on the U.S cities road dataset. We outperform the baselines across all datasets. We present the standard deviations in addition to the main results.

Fine Tuning Training Set	**(Pre-)Training Strategy**	**Node-mAP↑**	**Node-mAR↑**	**Edge-mAP↑**	**Edge-mAR↑**	**SMD ↓**	**Topo-Prec.↑**	**Topo-Rec.↑**
A) TL from roads (2D) to roads (2D)								
Agadez [20]	No Pretr. [18]	0.067±0.006	0.122±0.007	0.021±0.005	0.043±0.006	0.062±0.028	0.369±0.051	0.261±0.047
	Self-superv. [9]	0.083±0.010	0.156±0.011	0.030±0.005	0.071±0.007	0.030±0.005	0.471±0.0082	0.459±0.039
	Supervised	0.161±0.021	0.237±0.023	0.115±0.016	**0.177±0.017**	0.023±0.009	0.783±0.018	**0.711±0.039**
	Ours	**0.163±0.017**	**0.244±0.015**	**0.116±0.019**	0.172±0.021	**0.022±0.003**	**0.816±0.032**	0.614±0.036
Munich [20]	No Pretr. [18]	0.083±0.012	0.120±0.011	0.034±0.013	0.054±0.016	0.235±0.049	0.260±0.057	0.247±0.070
	Self-superv. [9]	0.088±0.021	0.145±0.033	0.060±0.015	0.097±0.023	0.155±0.032	0.339±0.035	0.384±0.075
	Supervised	0.277±0.022	0.336±0.025	0.207±0.027	0.272±0.031	0.091±0.038	0.682±0.037	**0.660±0.041**
	Ours	**0.285±0.015**	**0.344±0.011**	**0.224±0.030**	**0.277±0.031**	**0.090±0.043**	**0.726±0.078**	0.655±0.070
B) TL from roads (2D) to synthetic retinal vessels (2D)								
Synthetic OCTA [35]	No Pretr. [18]	0.273±0.003	0.375±0.003	0.140±0.002	0.339±0.003	0.005±0.002	0.181±0.004	0.948±0.004
	Self-superv. [9]	0.136±0.002	0.260±0.003	0.069±0.002	0.223±0.004	0.031±0.006	0.093±0.005	0.927±0.010
	Supervised	0.291±0.003	0.384±0.003	0.170±0.002	0.338±0.005	0.004±0.001	0.211±0.005	**0.957±0.007**
	Ours	**0.415±0.005**	**0.493±0.003**	**0.250±0.004**	**0.415±0.004**	**0.002±0.001**	**0.401±0.003**	0.890±0.007
C) TL from roads (2D) to real retinal vessels (2D)								
OCTA-500 [26]	No Pretr. [18]	0.189±0.005	0.282±0.007	0.108±0.004	0.169±0.006	0.017±0.002	0.737±0.007	0.634±0.010
	Self-superv. [9]	0.214±0.004	0.305±0.004	0.135±0.001	0.213±0.002	0.016±0.002	0.763±0.012	0.706±0.005
	Supervised	0.366±0.004	0.447±0.004	0.276±0.006	0.354±0.007	0.014±0.001	0.862±0.010	0.775±0.011
	Ours	**0.491±0.006**	**0.571±0.005**	**0.366±0.009**	**0.489±0.007**	**0.012±0.002**	**0.877±0.004**	**0.817±0.011**
D) TL from roads (2D) to brain vessels (3D)								
Synthetic MRI [40]	No Pretr. [18]	0.162±0.003	0.250±0.003	0.125±0.004	0.201±0.004	**0.013±0.000**	-	-
	Self-superv. [9]	0.162±0.003	0.252±0.003	0.120±0.004	0.193±0.004	0.014±0.000	-	-
	Supervised	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Ours	**0.356±0.003**	**0.450±0.002**	**0.221±0.003**	**0.322±0.003**	**0.013±0.000**	-	-
E) TL from roads (2D) to real whole-brain vessel data (3D)								
Microscopic images [47]	No Pretr. [18]	0.231±0.016	0.308±0.021	0.249±0.017	0.329±0.023	**0.017±0.000**	-	-
	Self-superv. [9]	0.344±0.026	0.404±0.029	0.363±0.026	0.425±0.030	**0.017±0.000**	-	-
	Supervised	⋆	⋆	⋆	⋆	⋆	⋆	⋆
	Ours	**0.483±0.037**	**0.535±0.039**	**0.523±0.041**	**0.566±0.043**	**0.017±0.000**	-	-

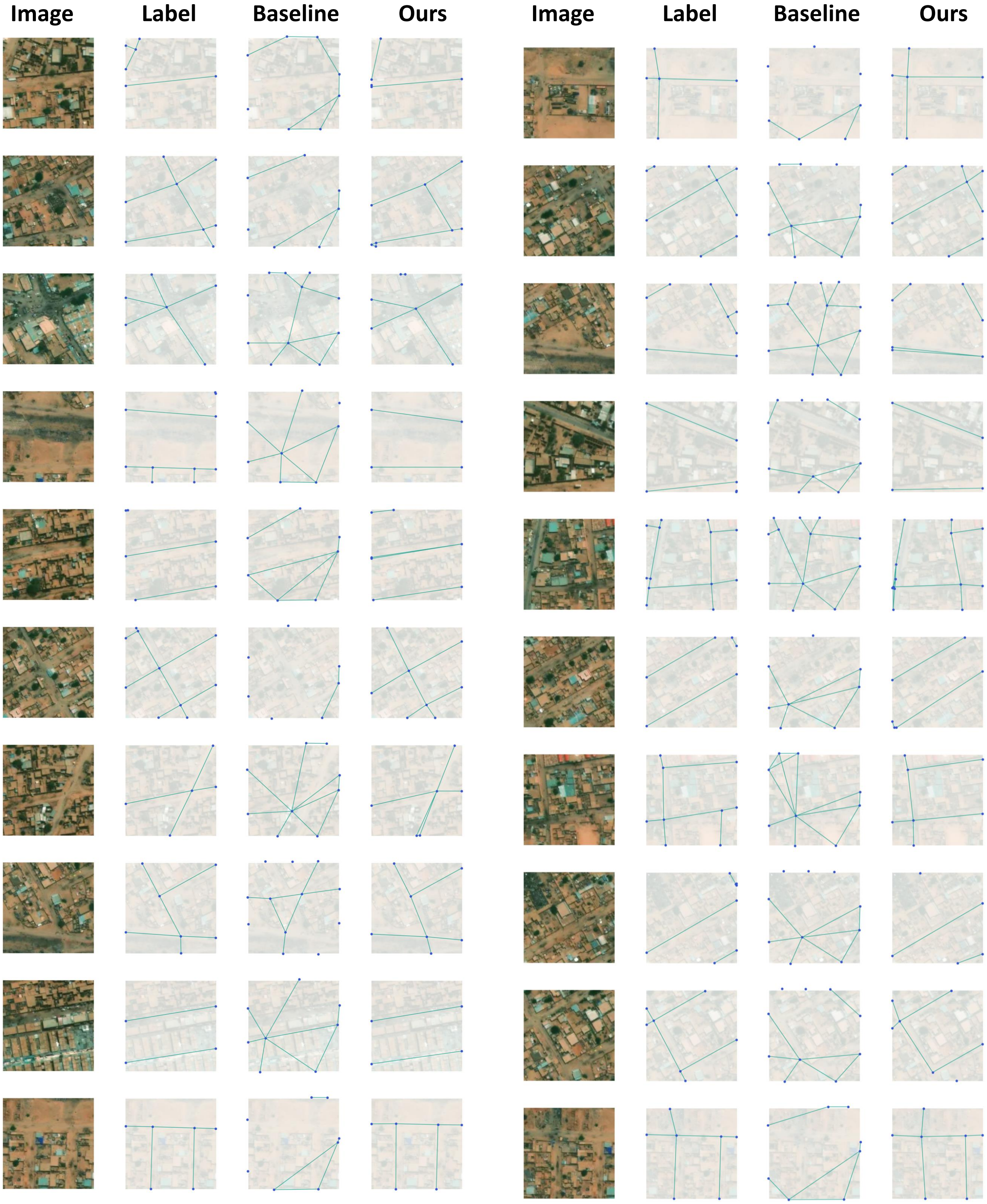

Figure 7. Qualitative results for the Agadez dataset. Two columns, from left to right: Image, ground truth graph, baseline, and our method. Our method consistently outperforms the baselines, which overpredict the edges and nodes for road data.

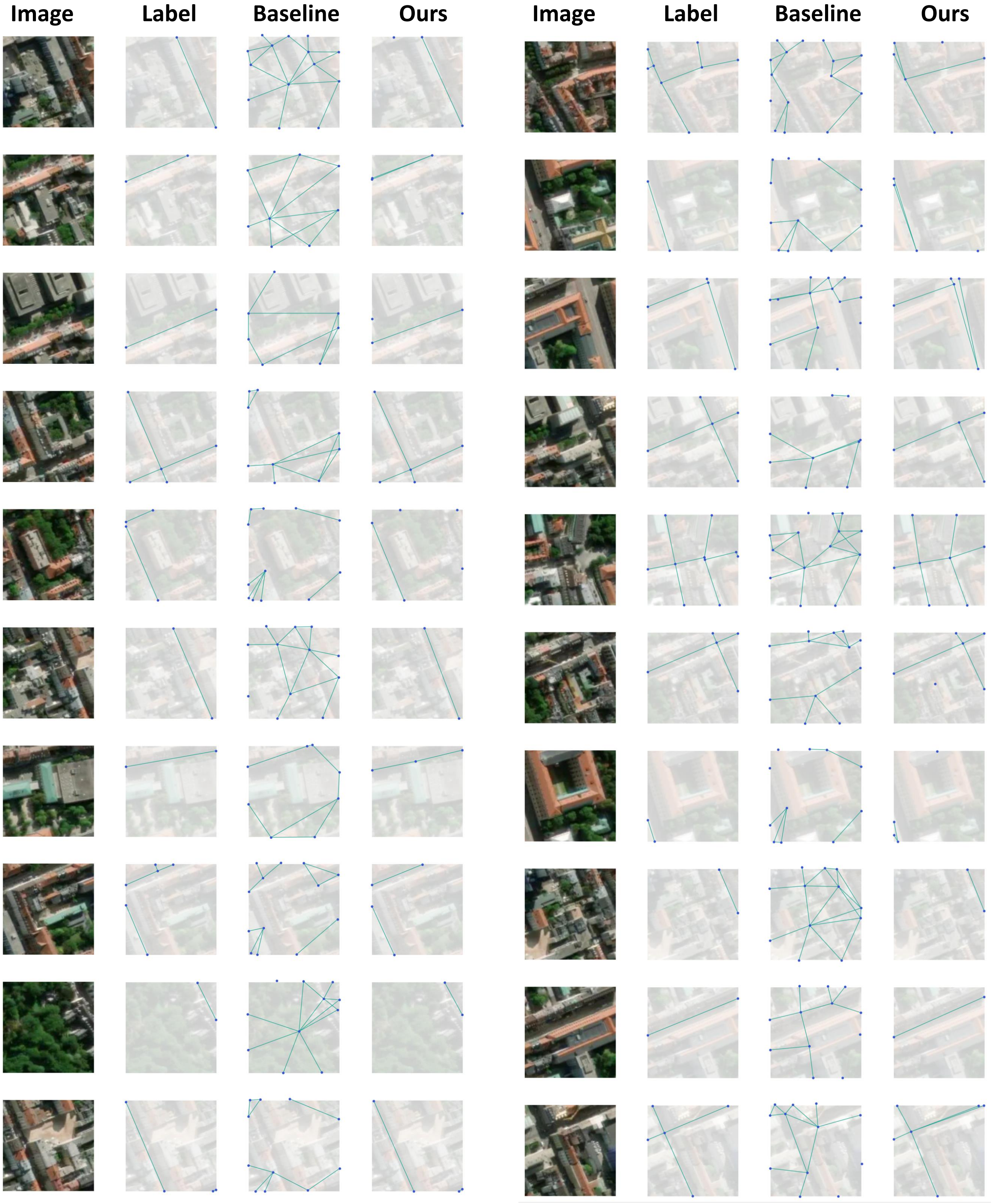

Figure 8. Qualitative results for the Munich dataset. Two columns, from left to right: Image, ground truth graph, baseline, and our method. Our method consistently outperforms the baselines, which overpredict the edges and nodes for road data.

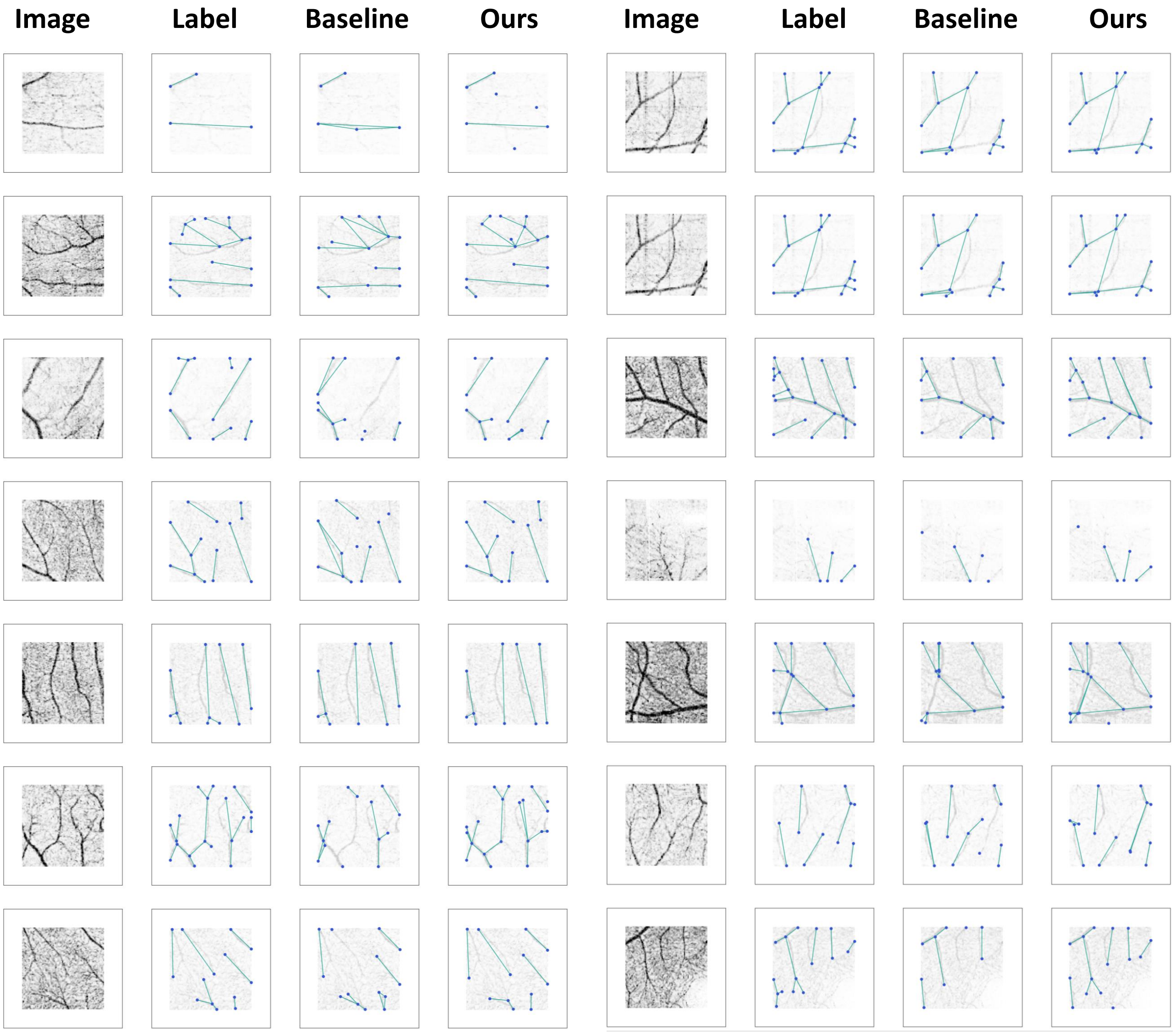

Figure 9. Qualitative results for the OCTA-500 dataset. Two columns, from left to right: Image, ground truth graph, baseline, and our method. Our method consistently outperforms the baselines, which underpredict the edges and nodes for the vessel data. It is important to note that the OCTA-500 dataset labels are on the large vessels. The graph annotations are not provided for all capillaries and are therefore not learned by the models either.

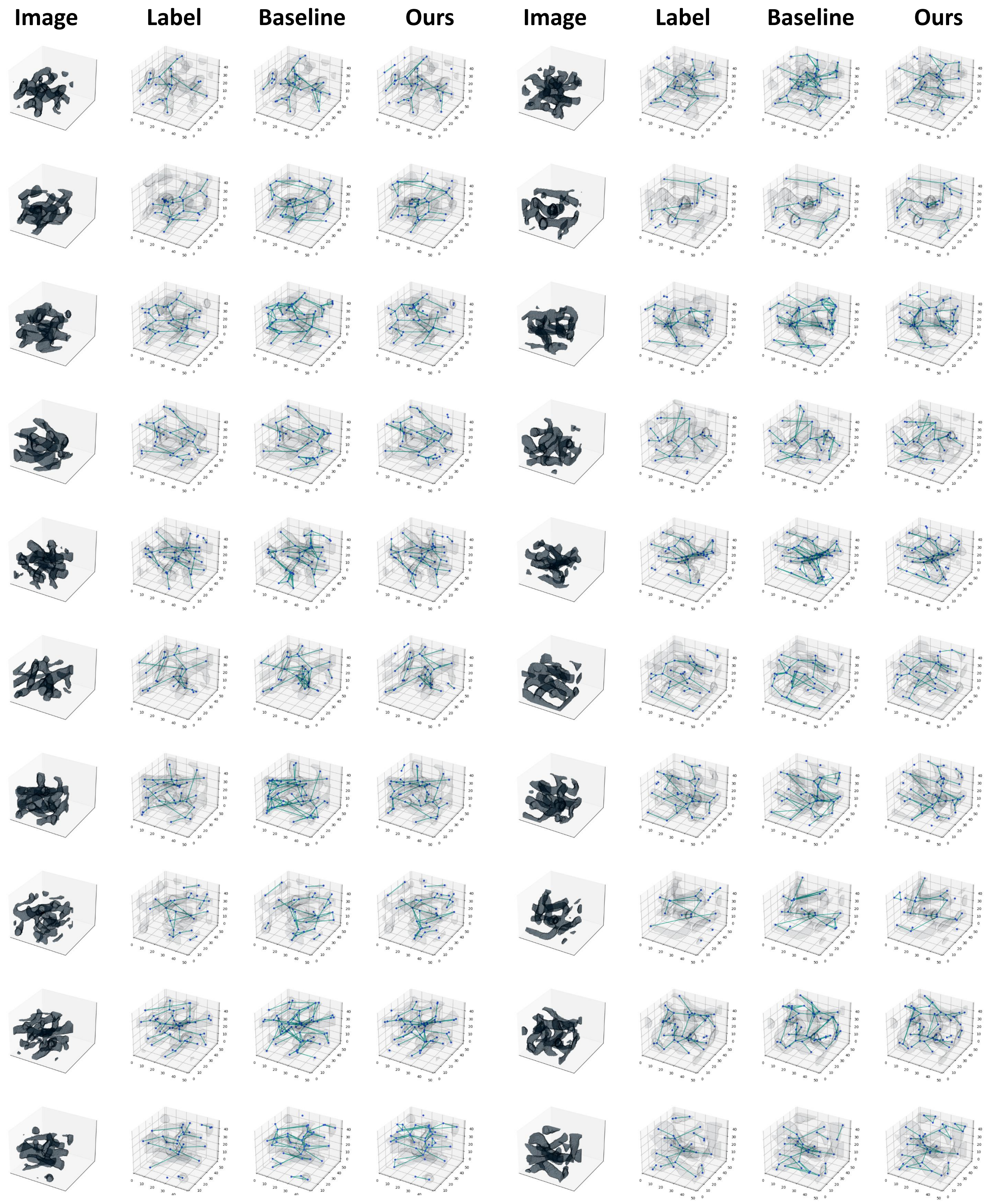

Figure 10. Qualitative results for the 3D whole brain vessel dataset. Two columns, from left to right: Image, ground truth graph, baseline, and our method. Our method consistently outperforms the baselines, which overpredict the edges for the 3D vessel data. Furthermore, the baseline often predicts implausible triangles between three nodes.

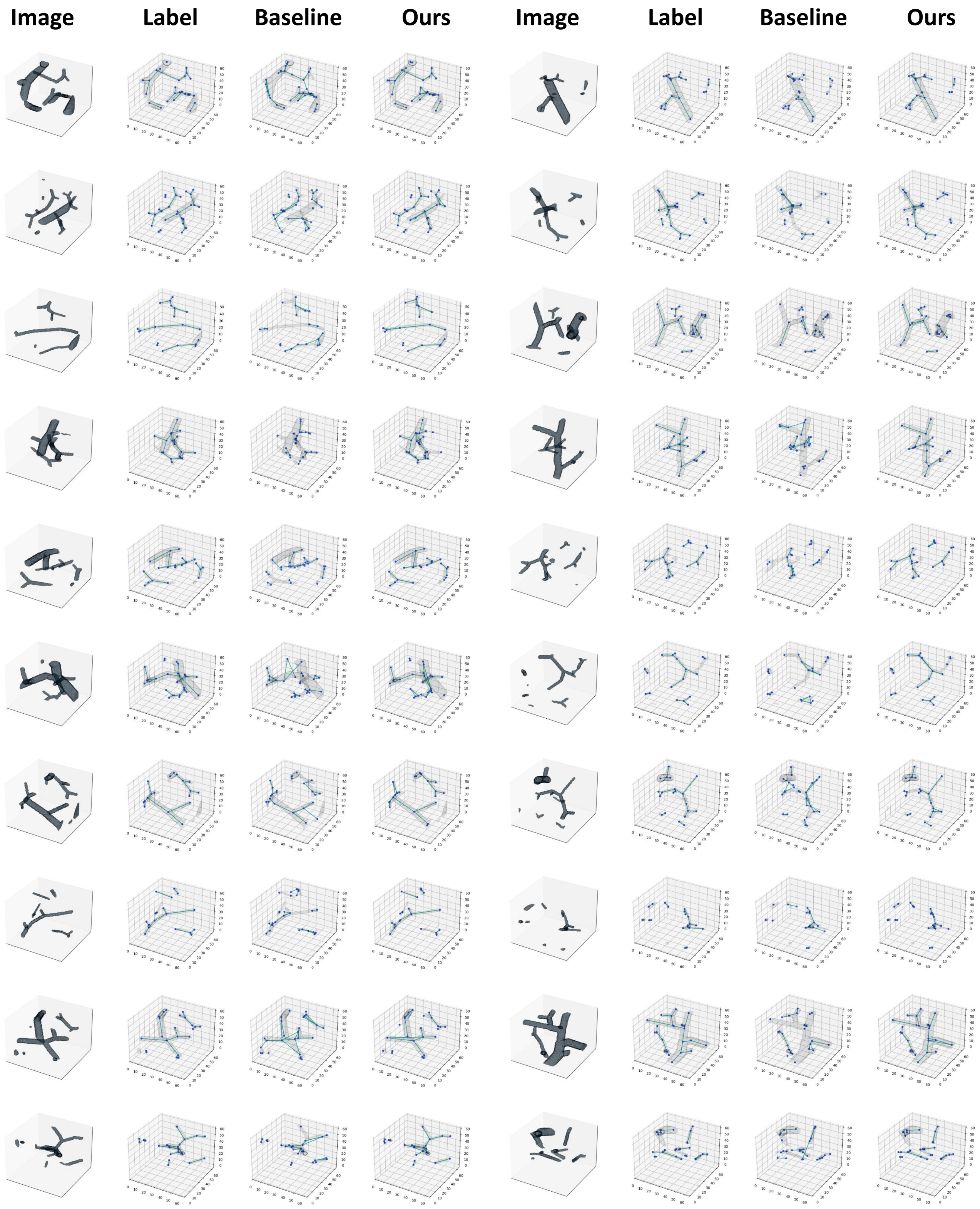

Figure 11. Qualitative results for the synthetic 3D vessel MRI dataset. Two columns, from left to right: Image, ground truth graph, baseline, and our method. Our method consistently outperforms the baselines, which overpredict the nodes for the 3D vessel data and underpredict edges.