Table 6: Results using UNet backbone on the ROADS dataset. We also perform the unpaired t-test (95% confidence interval) to determine the statistical significance. The statistically significant better performances are highlighed in bold

Dataset	Method	ECE (%)↓	Dice↑	clDice↑	ARI↑	VOI↓
ROADS	UNet	9.6491 ± 0.0049	0.7011 ± 0.0426	0.7918 ± 0.0679	0.7143 ± 0.0526	0.5832 ± 0.0345
	Prob.-UNet	8.4318 ± 0.0042	0.7194 ± 0.0418	0.8058 ± 0.0615	0.7350 ± 0.0494	0.5602 ± 0.0308
	PHiSeg	7.9331 ± 0.0038	0.7203 ± 0.0366	0.8113 ± 0.0521	0.7392 ± 0.0416	0.5559 ± 0.0295
	Hu et al.	7.8034 ± 0.0029	0.7275 ± 0.0361	0.8282 ± 0.0493	0.7314 ± 0.0391	0.5644 ± 0.0239
	Ours	**4.1442 ± 0.0031**	**0.7461 ± 0.0364**	**0.8496 ± 0.0455**	**0.7601 ± 0.0349**	**0.5463 ± 0.0218**

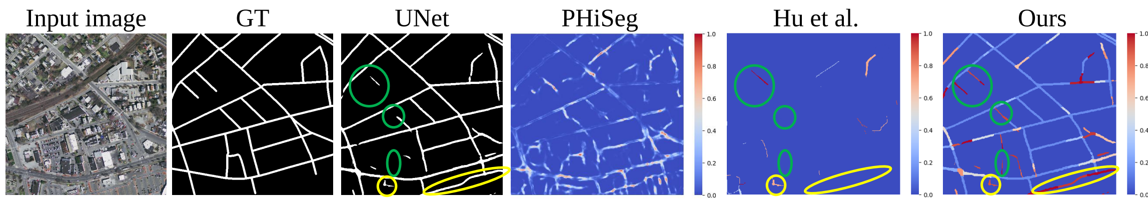

Figure 14: Qualitative results compared to the uncertainty baselines on the ROADS dataset. We show the uncertainty estimates in the form of a heatmap. Green highlights false negatives and yellow highlights false positives.

Table 7: Comparison on topological metrics Betti Number error and Betti Matching error for different datasets. All methods use UNet as the backbone. β_0^{err}, β_1^{err}, and β_2^{err} denote Betti Number error in 0-dim, 1-dim, and 2-dim respectively. μ_0^{err} and μ_1^{err} denote Betti Matching error in 0-dim and 1-dim respectively. The statistically significant better performances are highlighted in bold

Dataset	Method	$\beta_0^{\text{err}}\downarrow$	$\beta_1^{\text{err}}\downarrow$	$\beta_2^{\text{err}}\downarrow$	$\mu_0^{\text{err}}\downarrow$	$\mu_1^{\text{err}}\downarrow$
DRIVE	UNet	166.3154 ± 12.1065	9.3149 ± 4.2062	–	205.8312 ± 12.4496	28.9587 ± 5.3766
	Prob.-UNet	146.6373 ± 11.0831	7.8197 ± 3.1980	–	191.2790 ± 11.2324	24.7826 ± 4.9684
	PHiSeg	145.3777 ± 12.4873	7.1542 ± 3.6436	–	190.0528 ± 10.3376	24.0893 ± 4.1123
	Hu et al.	140.8317 ± 11.5502	6.3083 ± 2.2372	–	188.6573 ± 10.2403	23.7263 ± 4.3402
	Ours	**127.4041 ± 10.7344**	**4.6172 ± 2.6586**	–	**161.4536 ± 9.7017**	**20.6835 ± 3.4121**
ROSE	UNet	231.5081 ± 15.5573	9.8826 ± 1.6486	–	243.2775 ± 16.9274	14.8922 ± 1.6793
	Prob.-UNet	229.7987 ± 15.4307	9.0396 ± 1.6315	–	240.3295 ± 16.8371	14.0771 ± 2.0375
	PHiSeg	220.0644 ± 14.6356	7.8644 ± 2.0692	–	228.4348 ± 17.8907	11.0377 ± 1.9442
	Hu et al.	219.7530 ± 15.8446	7.6981 ± 1.5677	–	226.2989 ± 16.1992	10.6838 ± 1.8091
	Ours	**203.5791 ± 13.6467**	**5.0553 ± 1.4734**	–	**210.1763 ± 15.1485**	**8.6489 ± 1.5646**
ROADS	UNet	75.6666 ± 8.1079	25.5777 ± 7.4432	–	78.2291 ± 9.5055	30.5104 ± 6.6921
	Prob.-UNet	70.3564 ± 7.5929	24.3852 ± 7.0812	–	72.8129 ± 9.1638	29.8830 ± 6.1977
	PHiSeg	68.7237 ± 7.9177	24.1772 ± 6.5982	–	70.2788 ± 8.2474	28.9467 ± 5.9164
	Hu et al.	61.5167 ± 6.1625	23.5863 ± 5.3985	–	62.4951 ± 7.7601	26.2681 ± 5.8736
	Ours	**45.6735 ± 5.9286**	**17.2653 ± 4.6162**	–	**47.1429 ± 6.7905**	**23.1837 ± 5.4451**
PARSE	UNet	673.7016 ± 23.9541	79.5825 ± 10.9693	18.4316 ± 2.9432	–	–
	Prob.-UNet	620.1903 ± 22.0012	51.4995 ± 8.4096	16.7046 ± 2.2419	–	–
	PHiSeg	587.2137 ± 22.6801	45.9331 ± 8.7251	15.8529 ± 3.0218	–	–
	Hu et al.	555.9788 ± 23.5735	40.0707 ± 8.2376	13.9498 ± 2.2883	–	–
	Ours	**520.4991 ± 22.4327**	**33.0532 ± 7.8453**	**10.3831 ± 2.1035**	–	–

Table 8: Additional ablation study results of our method on the DRIVE dataset using UNet as the backbone. The best results (as reported in Table 1 of the main paper) are in bold. In the table, *Feature vector* denotes the length of the input feature vector to the GNN, while *Bounding box* denotes the size of the crop/bounding box centered on each structure. These hyperparameters are described in Sec 3.2 of the main paper. The $+1$ denotes the concatenation of the scalar persistence value

Feature vector	ECE (%)↓	clDice↑
$16+1$	4.9621 ± 0.0048	0.7906 ± 0.0394
$32+1$	**4.1633 ± 0.0043**	**0.7974 ± 0.0372**
$64+1$	4.1667 ± 0.0045	0.7972 ± 0.0367
Bounding box	**ECE (%)↓**	**clDice↑**
16×16	5.1085 ± 0.0047	0.7842 ± 0.0416
32×32	**4.1633 ± 0.0043**	**0.7974 ± 0.0372**
64×64	4.1689 ± 0.0042	0.7921 ± 0.0363