EndoGeoNet: Geometry-Arbitrated Thin-Branch Segmentation for Forest Navigation

Anonymous authors
Paper under double-blind review

Abstract

Thin branches are difficult to segment in forest scenes because they occupy few pixels and often share appearance with surrounding vegetation. We introduce EndoGeoNet, a monocular segmentation framework that augments a strong RGB backbone with geometry estimated from the same input image. A frozen monocular depth estimator provides a relative-depth cue, which is encoded by a lightweight multiscale geometry branch and fused at four decoder stages through Geo-Visual Arbitration Modules (GVAMs). Each GVAM learns stage-specific appearance–geometry weights, allowing the model to suppress unreliable depth while retaining useful geometric corrections. We also introduce the Forestry Sparse-Branch Dataset (FSB), a dedicated benchmark of 60 manually annotated full-HD images acquired across different locations and capture batches, together with topology- and component-aware evaluation. On FSB, EndoGeoNet improves the RGB baseline from 0.587 to 0.629 Dice and from 0.537 to 0.580 clDice, while maintaining near-identical performance on DIS-5K (0.899 versus 0.900 Dice). Replacing aligned depth with zero or shuffled depth reduces Dice by 2.19 and 2.63 points, respectively, showing that the model relies on spatially corresponding geometry. Including online depth estimation, the full pipeline runs at 7.74 FPS in FP32 and 16.74 FPS with mixed precision on an RTX 4090.

1 Introduction

Forestry robots and unmanned aerial vehicles (UAVs) need to perceive thin branches before local planning can avoid them. Leafless branches occupy few pixels, exhibit weak contrast, and resemble background vegetation. These properties connect branch perception to dichotomous image segmentation (DIS), where models must recover fine and complex foreground structures (Qin et al., 2022). Strong RGB models improve general DIS accuracy, yet appearance alone remains ambiguous when a branch and its background share texture, color, and contour statistics.

Geometry can resolve some of this ambiguity, but obtaining it is difficult. Direct depth sensors add payload, calibration, and power costs, and sparse ranging measurements may miss thin structures. A monocular estimator can instead derive relative geometry from the same RGB frame, yielding an endogenous cue without an external sensor. Such estimates are imperfect precisely around fine structures, however, so indiscriminate fusion can corrupt a strong appearance representation. The central challenge is therefore to exploit estimated geometry selectively while preventing its errors from dominating the segmentation representation.

EndoGeoNet addresses this challenge with a lightweight multiscale branch that encodes estimated depth and a Geo-Visual Arbitration Module (GVAM) that fuses geometry with appearance at four decoder stages. GVAM globally summarizes the aligned features and predicts stage-specific scalar weights for a convex appearance–geometry combination. The weights allow geometry to act as a small corrective signal when useful and to be suppressed when unreliable. We use a frozen Depth Anything V2 Small (DAv2-S) model as the Ge-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

ometry Perception Sub-network (GPS-Net) (Yang et al., 2024). The method centers on the multiscale representation and adaptive fusion of this endogenous geometric cue.

EndoGeoNet improves overlap and centerline recovery on sparse forestry branches while leaving general DIS accuracy nearly unchanged. Zero and shuffled-depth ablations show that the improvement depends on spatial alignment, while the learned weights reveal how much geometry contributes at each decoder stage. The paper makes three contributions:

- EndoGeoNet derives a relative geometric cue from the input RGB image and represents it with a lightweight multiscale pathway tailored to thin, low-contrast obstacles.
- GVAM performs stage-wise arbitration between appearance and geometry, adapting the contribution of the estimated depth cue at each decoder scale while preserving the strong RGB representation.
- FSB provides 60 manually annotated full-HD images acquired across different locations and capture batches. Its evaluation protocol covers region overlap, centerline continuity, navigation-corridor recall, and missed branch components.

2 Related Work

Automatic dichotomous image segmentation. DIS addresses foreground objects whose thin, complex structures are poorly represented by conventional region-oriented segmentation metrics and architectures. IS-Net formalized the task and introduced DIS-5K using a U^2-Net backbone (Qin et al., 2022; 2020). Subsequent methods strengthen RGB representations through complementary mechanisms. InSPyReNet refines predictions across an image pyramid, FP-DIS introduces frequency priors, and UDUN separates trunk and structural reasoning (Kim et al., 2022; Zhou et al., 2023; Pei et al., 2023). DC-Net divides the prediction problem across multiple encoders, while BADIS, BiRefNet, and MVANet emphasize boundaries, bilateral references, and multi-view aggregation, respectively (Zhu et al., 2025; Tang et al., 2024; Zheng et al., 2024; Yu et al., 2024). DiffDIS further exploits generative priors from a pretrained diffusion model (Yu et al., 2025). These methods achieve strong general DIS accuracy, but their appearance features remain ambiguous when sparse branches resemble surrounding vegetation.

Foundation and promptable segmentation models. Foundation models broaden segmentation through large-scale pretraining and prompt-conditioned mask generation. SAM established a general promptable interface, followed by adaptations based on conditional tuning, BA-SAM's scalable Bias-Mode Attention Mask, and prompt adapters for more efficient specialization (Kirillov et al., 2023; Xiao et al., 2024; Song et al., 2024; Xie et al., 2024). DIS-SAM couples SAM with a DIS network, while LawDIS introduces language-window control for dichotomous segmentation (Liu et al., 2025; Yan et al., 2025). Other work repurposes diffusion features for dense perception or combines complementary foundation-model encoders for automatic high-resolution segmentation (Xu et al., 2025; Xiong et al., 2025). Promptable methods provide useful control when an object cue is available, whereas forest navigation requires dense predictions without user interaction at inference time. EndoGeoNet adopts the automatic SAM2-UNeXT appearance pathway and augments it with geometry inferred from the same frame.

Geometry-aware segmentation. RGB-D segmentation uses measured depth to resolve appearance ambiguity. MAGNet and CPNet develop multiscale and progressive cross-modal fusion for RGB-D saliency, while DACOD and RISNet exploit depth for camouflaged or concealed objects (Zhong et al., 2024; Hu et al., 2024; Wang et al., 2023; 2024). These approaches demonstrate the value of geometry but assume a depth sensor, which adds payload, calibration, and power requirements to a small aerial platform. For DIS, PDFNet estimates pseudo-depth from RGB and combines patch-depth fusion with an integrity-prior loss (Liu et al., 2026). EndoGeoNet also derives geometry monocularly but treats the estimate as an imperfect auxiliary cue. A lightweight geometry encoder constructs multiscale features, and GVAM assigns an independent appearance–geometry balance at each decoder stage.

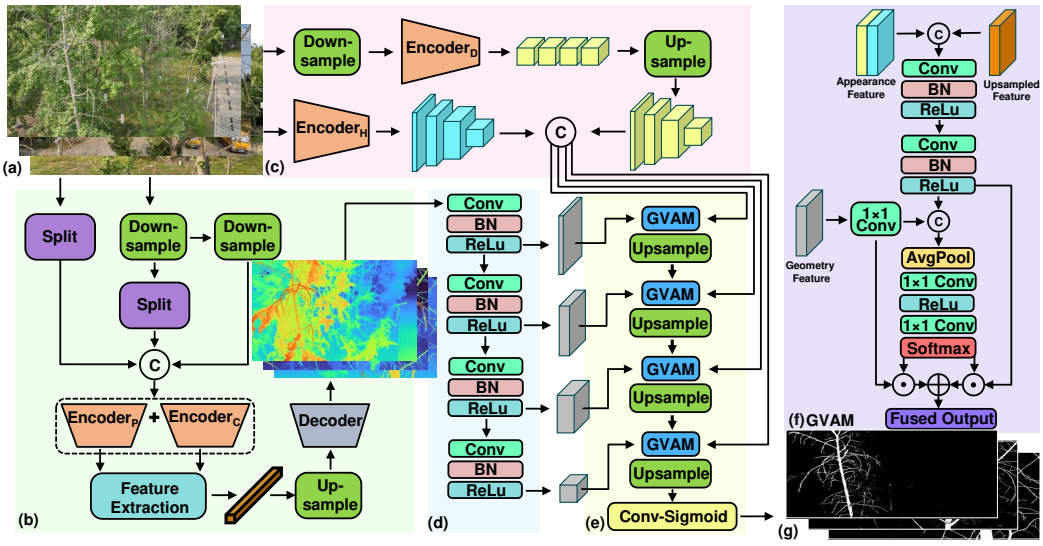


Figure 1: EndoGeoNet uses a frozen DAv2-S model as GPS-Net to derive a depth cue from the input RGB image. A lightweight geometry pathway encodes the cue, and GVAM fuses it with appearance features at four decoder stages. Training updates the geometry encoder and GVAM while keeping the appearance encoder and decoder fixed.

Ablations with zero and shuffled depth isolate the contribution of spatially corresponding geometry.

Monocular depth estimation. Modern monocular depth models recover dense relative geometry from a single image. Depth Anything V2 provides multiple model scales and preserves fine spatial detail while supporting efficient inference (Yang et al., 2024). EndoGeoNet uses the frozen small variant to generate relative-depth cues; segmentation training does not update the estimator or require metric-depth supervision. The research problem is therefore robust use of an imperfect estimated cue, especially near thin structures where monocular depth can be unreliable.

3 Method

3.1 Endogenous geometry and multiscale encoding

Given an RGB image $I \in \mathbb{R}^{H\times W\times 3}$, a frozen GPS-Net $\mathcal{P}$ estimates a relative depth map $D = \mathcal{P}(I)$. We use DAv2-S with an input size of 518 and FP16 inference. The output is resized to the segmentation resolution, oriented so larger values indicate farther regions, and normalized per image:

$$\bar{D} = \frac{D - \min(D)}{\max(D) - \min(D) + \epsilon}. \tag{1}$$

GPS-Net remains fixed during segmentation training. A four-block CNN geometry encoder produces $\mathbb{G} = \{G^k\}_{k=1}^4$ with channels $\{32, 64, 128, 128\}$ and resolutions $\{1/2, 1/4, 1/8, 1/16\}$:

$$G^k = \mathrm{ReLU}\big(\mathrm{BN}_k\left(\mathrm{Conv}^{3\times 3}_{s=2}(G^{k-1})\right)\big), \quad G^0 = \bar{D}. \tag{2}$$

The appearance pathway follows the dual-stream Hiera–DINOv2 encoder of SAM2-UNeXT (Xiong et al., 2025; Oquab et al., 2024). It yields appearance features $\mathbb{A} = \{A^k\}_{k=1}^4$. The appearance encoder and decoder are initialized with the SAM2-UNeXT RGB weights and remain fixed during training. Optimization updates only the geometry encoder and GVAM.

3.2 Geo-Visual Arbitration Module

At decoder stage k, the base decoder combines the upsampled preceding feature with the appearance skip to obtain F^k. A 1×1 convolution aligns G^k with its channel and spatial dimensions:

$$\widetilde{G}^k = \mathrm{Resize}(\mathcal{C}^k_{\mathrm{align}}(G^k), \mathrm{shape}(F^k)). \tag{3}$$

GVAM concatenates F^k and $\widetilde{G}^k$, applies global average pooling, and maps the descriptor through a two-layer bottleneck MLP. A Softmax produces stage-level scalar weights:

$$v^k = \mathrm{GAP}([F^k, \widetilde{G}^k]), \tag{4}$$

$$(w^k_A, w^k_G) = \mathrm{Softmax}(W^k_2\,\mathrm{ReLU}(W^k_1 v^k + b^k_1) + b^k_2), \tag{5}$$

$$U^k = w^k_A F^k + w^k_G \widetilde{G}^k. \tag{6}$$

Thus $w^k_A + w^k_G = 1$. The weights vary across images and stages but remain spatially global within a stage. This keeps GVAM compact (352,840 trainable parameters together with the geometry encoder) and makes the arbitration interpretable. We initialize the MLP output logits with an appearance-favoring prior, which provides nonzero gradients for both modalities from the first update.

3.3 Training objective

We optimize the geometry encoder and four GVAM instances with

$$\mathcal{L} = \mathcal{L}_{\mathrm{wBCE}} + \mathcal{L}_{\mathrm{wIoU}}, \tag{7}$$

where both terms use the boundary-sensitive weight map $W = 1 + 5|\mathrm{AvgPool}_{31}(Y) - Y|$ for ground-truth mask Y. The weighted IoU term uses Laplace smoothing of one. All appearance parameters and their batch-normalization statistics remain frozen.

4 Experiments

Datasets. We train EndoGeoNet on DIS-TR (3,000 images), use DIS-VD (470 images) for model selection, and evaluate it on the four DIS test splits (500 images each).

FSB is a dedicated thin-branch benchmark containing 60 manually annotated 1920×1080 forestry images acquired across different locations and capture batches. The annotation target is visible woody foreground that may act as an obstacle, including connected tree structures when discernible at image resolution. Masks are drawn with polygon and brush tools and inspected as image overlays. For terminal twigs, annotators preserve visible centerline continuity instead of smoothing fine boundaries. Branch pixels occupy $9.30\% \pm 4.83\%$ of an image, and masks contain 589.4 ± 551.3 connected components per image, reflecting the sparsity and fragmentation of the target. Figure 2 illustrates the variation in branch density and background clutter. A second annotator independently labels five representative images; the minimum agreement is 0.961 IoU, 0.980 Dice, 0.951 clDice, and 0.945 boundary F-measure. Appendix B gives further dataset statistics.

Method development, training, and model selection use DIS-5K. FSB serves only as an evaluation benchmark, and none of its images are used for target-domain fine-tuning.

Training details. Images, masks, and depth cues are resized to 1024×1024. Training uses random horizontal and vertical flips, AdamW with a learning rate of 2×10^{-4} and weight decay of 5×10^{-4}, cosine decay, batch size one, and automatic mixed precision. We train for eight epochs and use mean DIS-VD Dice at threshold 0.5 for model selection.

Metrics and efficiency. We report S_α, weighted and maximum F-measure, E_ϕ, MAE, IoU, Dice, and clDice. For forestry data, skeleton F1 matches predicted and ground-truth centerlines within a two-pixel tolerance. Central-corridor recall divides the correctly predicted foreground pixels inside the middle 50% of the image width by the ground-truth foreground pixels in the same region. Component miss rate is the fraction of ground-truth connected



Figure 2: Four representative FSB images and manual pixel masks. The target includes visible woody foreground and connected tree structure, with terminal-twig centerline continuity preserved where discernible.

Table 1: Five-split macro average on DIS-5K without test-time augmentation. Higher is better except MAE.

Method	S_α	F_β^w	$F_\beta^{\max}$	E_ϕ	MAE↓	IoU	Dice	clDice
SAM2-UNeXT	0.912	0.871	0.909	0.940	0.032	0.839	0.900	0.895
DiffDIS	0.916	0.847	0.916	0.919	0.036	0.834	0.896	0.902
PDFNet	0.891	0.839	0.890	0.924	0.041	0.778	0.862	0.859
BiRefNet	0.900	0.859	0.895	0.935	0.035	0.820	0.887	0.889
MVANet	0.909	0.865	0.907	0.937	0.034	0.825	0.891	0.893
BADIS	0.863	0.794	0.851	0.908	0.050	0.748	0.838	0.824
DC-Net	0.840	0.759	0.819	0.880	0.063	0.711	0.808	0.807
EndoGeoNet (DAv2-S)	0.911	0.867	0.909	0.940	0.032	0.839	0.899	0.897

components of at least 20 pixels in that corridor with no predicted overlap. Predictions are thresholded at 0.5 and resized to the native ground-truth resolution. Online latency includes depth estimation and segmentation, excludes disk I/O, and uses batch size one on an RTX 4090. FP32 is used for the architecture comparison, and AMP characterizes mixed-precision deployment.

5 Results

5.1 Public and target-domain segmentation

Table 1 compares EndoGeoNet with seven recent DIS methods. DiffDIS obtains the highest S_α, maximum F-measure, and clDice, whereas EndoGeoNet matches the best E_ϕ, MAE, and IoU. Relative to its SAM2-UNeXT appearance backbone, EndoGeoNet changes Dice from 0.900 to 0.899 and clDice from 0.895 to 0.897. The geometry pathway therefore preserves general DIS performance, but it does not produce a consistent aggregate gain on this broad benchmark.

The target-domain results differ markedly from the DIS-5K averages. EndoGeoNet ranks first on every reported FSB metric in Table 2. Relative to SAM2-UNeXT, it improves IoU by 0.036, Dice by 0.042, and clDice by 0.043, while reducing the component miss rate from 0.563 to 0.498. The simultaneous improvement in overlap, centerline agreement, and component detection indicates that the gain extends beyond large trunk regions to sparse branch structures that are easily omitted by an appearance-only model.

The other DIS methods transfer less effectively to FSB. Their Dice scores range from 0.074 to 0.306, and their component miss rates remain between 0.848 and 0.952. This gap reflects the target-domain difficulty: thin woody structures occupy a small image area, share local appearance with surrounding vegetation, and often coexist with more salient objects such as vehicles, poles, and road boundaries.

Table 2: Results on the 60-image Forestry Sparse-Branch Dataset (FSB). Higher is better except MAE and component miss rate. All methods are evaluated without target-domain fine-tuning or test-time augmentation.

Method	S_α	MAE↓	IoU	Dice	clDice	Miss↓
SAM2-UNeXT	0.724	0.071	0.449	0.587	0.537	0.563
DC-Net	0.560	0.094	0.195	0.294	0.251	0.867
DiffDIS	0.532	0.101	0.103	0.152	0.126	0.912
PDFNet	0.518	0.091	0.129	0.198	0.173	0.877
BiRefNet	0.470	0.094	0.046	0.074	0.047	0.952
MVANet	0.565	0.090	0.208	0.306	0.279	0.848
BADIS	0.555	0.085	0.184	0.275	0.219	0.871
EndoGeoNet	0.744	0.070	0.485	0.629	0.580	0.498

Table 3: Effect of replacing aligned depth with zero, shuffled, or absent depth at inference. Paired bootstrap confidence intervals for the Dice and clDice differences exclude zero.

Depth condition	IoU	Dice	clDice
Correctly aligned	0.509	0.649	0.596
Zero	0.487	0.627	0.575
Shuffled	0.483	0.623	0.572
Absent	0.484	0.625	0.574

Figure 3 provides a scene-level view of these errors. Several appearance-based methods produce sparse or fragmented branch responses and sometimes concentrate on high-contrast background objects. SAM2-UNeXT retains more woody structure than the other RGB competitors in several scenes, but its predictions still omit many terminal branches. PDFNet is an important geometry-aware comparison because it also uses estimated depth, yet its responses remain weak or shift toward salient background regions in these examples. This behavior shows that an estimated depth map is not sufficient unless its contribution is integrated selectively with appearance features.

EndoGeoNet produces denser and more continuous branch responses across most of the seven scenes. This visual pattern agrees with its higher clDice and lower component miss rate on the complete FSB benchmark. Errors remain around background foliage and vehicles, where the estimated geometry and visual structure can resemble the target. The qualitative results therefore support the benefit of geo-visual arbitration while also exposing the remaining foreground-background ambiguity.

5.2 Dependence on spatially aligned geometry

At inference, we replace the predicted depth while leaving the segmentation network unchanged. Aligned depth improves Dice by 0.0219 over zero depth and 0.0263 over shuffled depth; the corresponding 95% paired-bootstrap intervals are $[0.0130, 0.0331]$ and $[0.0165, 0.0384]$. The clDice gains are 0.0212 $[0.0138, 0.0303]$ and 0.0246 $[0.0169, 0.0344]$. Each interval uses 100,000 image-level bootstrap samples from paired FSB predictions. The consistent decrease with zero and shuffled depth shows that GVAM benefits from spatial correspondence rather than depth statistics alone.

5.3 Fusion design

GVAM outperforms both concatenation controls on every reported FSB metric (Table 4). Compared with projection-based concatenation, it improves Dice by 0.0108 and clDice by 0.0147 while using 76% fewer trainable fusion-and-geometry parameters. Residual concatenation performs worse still. These results support learned stage-wise modality arbitration over fixed feature aggregation for sparse branches.

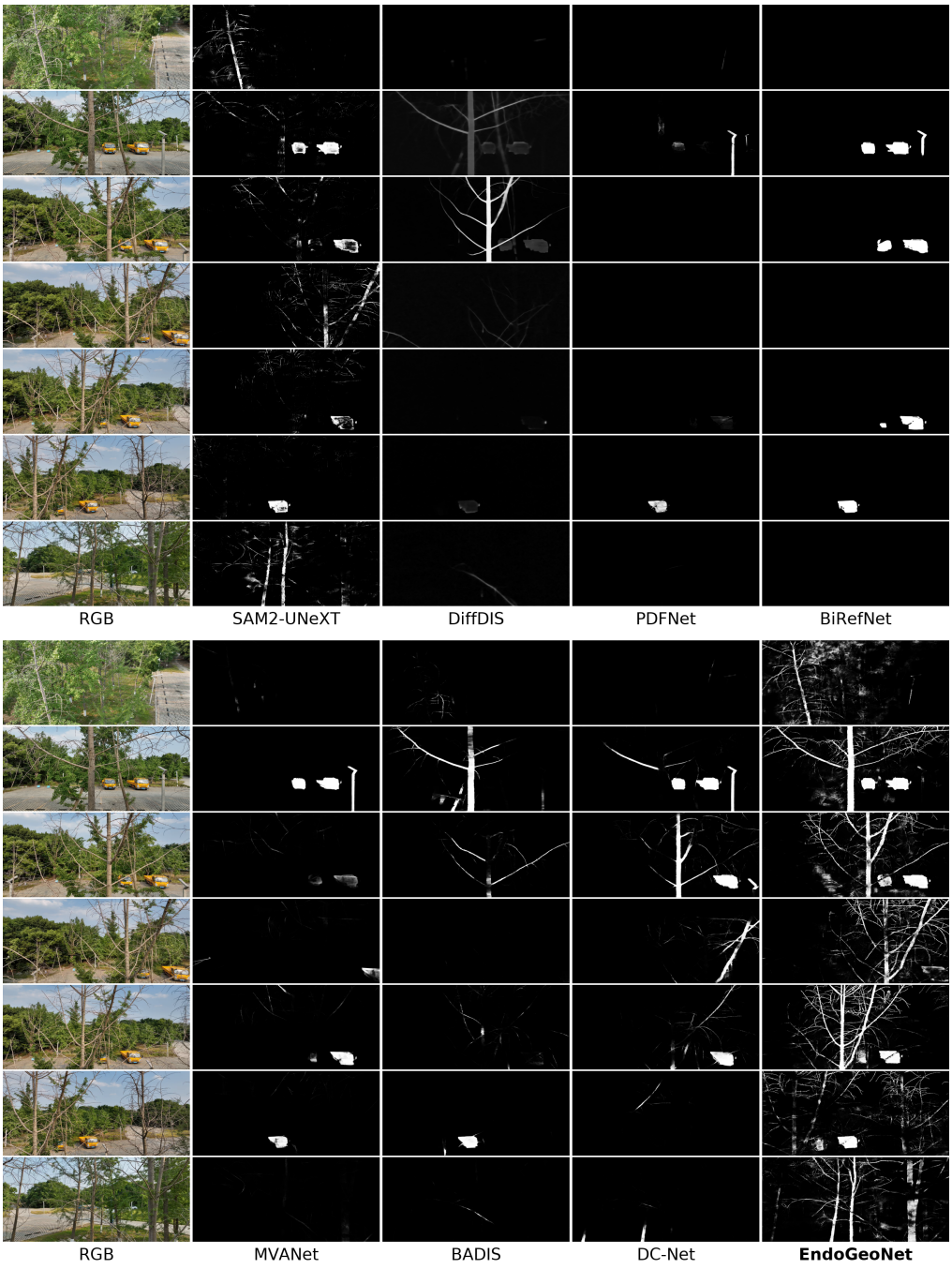


Figure 3: Qualitative comparison on seven representative FSB scenes. The upper block compares RGB inputs with SAM2-UNeXT, DiffDIS, PDFNet, and BiRefNet. The lower block compares the same inputs with MVANet, BADIS, DC-Net, and EndoGeoNet. Prediction intensity denotes foreground probability.

5.4 Online efficiency

The complete FP32 EndoGeoNet pipeline adds 18.0 ms to the RGB backbone and operates at 7.74 FPS with online geometry estimation. Mixed-precision execution reaches 16.74 FPS. These measurements characterize workstation-GPU performance; deployment speed on a lightweight onboard computer requires separate evaluation.

Table 4: Comparison of GVAM with direct concatenation alternatives. All variants share the appearance backbone, DAv2-S depth cues, training data, and optimization schedule. Trainable parameters include the geometry encoder and fusion modules.

Fusion	Trainable params	IoU	Dice	clDice	F_β^w
GVAM	0.353M	0.5088	0.6492	0.5963	0.5732
Concat + projection	1.467M	0.5048	0.6384	0.5816	0.5614
Residual concat	1.467M	0.4919	0.6222	0.5612	0.5446

Table 5: Full online efficiency on one RTX 4090 at batch size one and 1024×1024 segmentation input. Depth estimation and segmentation are both included; I/O is excluded.

Model	Params	Latency	FPS	Memory
SAM2-UNeXT, FP32	522.1M	111.2 ms	8.99	2906 MiB
DAv2-S + GVAM, FP32	547.3M	129.2 ms	7.74	3247 MiB
DAv2-S + GVAM, AMP	547.3M	59.7 ms	16.74	3247 MiB

6 Limitations and Scope

FSB is independent of method development and spans different locations and capture batches. Its 60 images nevertheless cover a limited range of forest environments. Larger evaluations across sites, seasons, sensors, illumination conditions, and vegetation types are needed to establish broader forestry generalization.

7 Conclusion

EndoGeoNet combines appearance with monocularly derived geometry through a lightweight encoder and stage-wise GVAM. FSB supplies manually annotated full-HD images and a topology-aware evaluation protocol for sparse forestry branches. EndoGeoNet achieves 0.899 Dice on DIS-5K and improves SAM2-UNeXT by 4.2 Dice points and 4.3 clDice points on FSB. Depth-input ablations confirm that these gains depend on spatially aligned geometry. The complete pipeline runs at 7.74 FPS in FP32 and 16.74 FPS with mixed precision on an RTX 4090. Broader domain generalization and closed-loop navigation remain open problems.

AI use statement

Generative AI tools were used only to assist with language editing, including grammar correction, wording refinement, and improvement of readability.

References

Xihang Hu, Fuming Sun, Jing Sun, Fasheng Wang, and Haojie Li. Cross-modal fusion and progressive decoding network for RGB-D salient object detection. International Journal of Computer Vision, 132(8):3067–3085, 2024. doi: 10.1007/s11263-024-02020-y.

Taehun Kim, Kunhee Kim, Joonyeong Lee, Dongmin Cha, Jiho Lee, and Daijin Kim. Revisiting image pyramid structure for high resolution salient object detection. In Computer Vision – ACCV 2022, pp. 257–273, 2022. doi: 10.1007/978-3-031-26293-7_16.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In IEEE/CVF International Conference on Computer Vision, pp. 3992–4003, 2023. doi: 10.1109/ICCV51070.2023.00371.

Xianjie Liu, Keren Fu, Yao Jiang, and Qijun Zhao. Promoting segment anything model towards highly accurate dichotomous image segmentation. In IEEE International Conference on Multimedia and Expo, pp. 1–6, 2025. doi: 10.1109/ICME59968.2025.11208915.

Xianjie Liu, Keren Fu, and Qijun Zhao. High-precision dichotomous image segmentation via depth integrity-prior and fine-grained patch strategy. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 6357–6366, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Jialun Pei, Zhangjun Zhou, Yueming Jin, He Tang, and Pheng-Ann Heng. Unite-divide-unite: Joint boosting trunk and structure for high-accuracy dichotomous image segmentation. In ACM International Conference on Multimedia, pp. 2139–2147, 2023. doi: 10.1145/3581783.3611811.

Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R. Zaiane, and Martin Jagersand. U^2-Net: Going deeper with nested U-structure for salient object detection. Pattern Recognition, 106:107404, 2020. doi: 10.1016/j.patcog.2020.107404.

Xuebin Qin, Hang Dai, Xiaobin Hu, Deng-Ping Fan, Ling Shao, and Luc Van Gool. Highly accurate dichotomous image segmentation. In European Conference on Computer Vision, pp. 38–56, 2022. doi: 10.1007/978-3-031-19797-0_3.

Yiran Song, Qianyu Zhou, Xiangtai Li, Deng-Ping Fan, Xuequan Lu, and Lizhuang Ma. BA-SAM: Scalable bias-mode attention mask for segment anything model. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3162–3173, 2024. doi: 10.1109/CVPR52733.2024.00305.

Haonan Tang, Shuhan Chen, Yang Liu, Shiyu Wang, Zeyu Chen, and Xuelong Hu. Boundary-aware dichotomous image segmentation. The Visual Computer, 40(12):9051–9062, 2024. doi: 10.1007/s00371-024-03295-5.

Liqiong Wang, Jinyu Yang, Yanfu Zhang, Fangyi Wang, and Feng Zheng. Depth-aware concealed crop detection in dense agricultural scenes. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 17201–17211, 2024. doi: 10.1109/CVPR52733.2024.01628.

Qingwei Wang, Jinyu Yang, Xiaosheng Yu, Fangyi Wang, Peng Chen, and Feng Zheng. Depth-aided camouflaged object detection. In ACM International Conference on Multimedia, pp. 3297–3306, 2023. doi: 10.1145/3581783.3611874.

Aoran Xiao, Weihao Xuan, Heli Qi, Yun Xing, Ruijie Ren, Xiaoqin Zhang, Ling Shao, and Shijian Lu. CAT-SAM: Conditional tuning for few-shot adaptation of segment anything model. In European Conference on Computer Vision, pp. 189–206, 2024. doi: 10.1007/978-3-031-73661-2_11.

Zhaozhi Xie, Bochen Guan, Weihao Jiang, Muyang Yi, Yue Ding, Hongtao Lu, and Lei Zhang. PA-SAM: Prompt adapter SAM for high-quality image segmentation. In IEEE International Conference on Multimedia and Expo, pp. 1–6, 2024. doi: 10.1109/ICME57554.2024.10687602.

Xinyu Xiong, Zihuang Wu, Lei Zhang, Lei Lu, Ming Li, and Guanbin Li. SAM2-UNeXT: An improved high-resolution baseline for adapting foundation models to downstream segmentation tasks, 2025. arXiv:2508.03566.

Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. What matters when repurposing diffusion models for general dense perception tasks? In International Conference on Learning Representations, 2025. URL https://openreview.net/forum?id=BgYbk6ZmeX.

Xinyu Yan, Meijun Sun, Ge-Peng Ji, Fahad Shahbaz Khan, Salman Khan, and Deng-Ping Fan. LawDIS: Language-window-based controllable dichotomous image segmentation. In IEEE/CVF International Conference on Computer Vision, pp. 23902–23911, 2025. doi: 10.1109/ICCV51701.2025.02216.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything V2. In Advances in Neural Information Processing Systems, volume 37, pp. 21875–21911, 2024. doi: 10.52202/079017-0688.

Qian Yu, Xiaoqi Zhao, Youwei Pang, Lihe Zhang, and Huchuan Lu. Multi-view aggregation network for dichotomous image segmentation. In IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3921–3930, 2024. doi: 10.1109/CVPR52733.2024.00376.

Qian Yu, Peng-Tao Jiang, Hao Zhang, Jinwei Chen, Bo Li, Lihe Zhang, and Huchuan Lu. High-precision dichotomous image segmentation via probing diffusion capacity. In International Conference on Learning Representations, 2025.

Peng Zheng, Dehong Gao, Deng-Ping Fan, Li Liu, Jorma Laaksonen, Wanli Ouyang, and Nicu Sebe. Bilateral reference for high-resolution dichotomous image segmentation. CAAI Artificial Intelligence Research, 3:9150038, 2024. doi: 10.26599/air.2024.9150038.

Mingyu Zhong, Jing Sun, Peng Ren, Fasheng Wang, and Fuming Sun. MAGNet: Multi-scale awareness and global fusion network for RGB-D salient object detection. Knowledge-Based Systems, 299:112126, 2024. doi: 10.1016/j.knosys.2024.112126.

Yan Zhou, Bo Dong, Yuanfeng Wu, Wentao Zhu, Geng Chen, and Yanning Zhang. Dichotomous image segmentation with frequency priors. In International Joint Conference on Artificial Intelligence, pp. 1822–1830, 2023. doi: 10.24963/ijcai.2023/202.

Jiayi Zhu, Xuebin Qin, and Abdulmotaleb Elsaddik. DC-Net: Divide-and-conquer for salient object detection. Pattern Recognition, 157:110903, 2025. doi: 10.1016/j.patcog.2024.110903.

A Per-split DIS-5K results

Table 6: EndoGeoNet with DAv2-S and GVAM on DIS-5K. The Average row is the unweighted macro average.

Split	S_α	F_β^w	$F_\beta^{\max}$	E_ϕ	MAE↓	IoU	Dice	clDice
DIS-VD	0.9105	0.8658	0.9047	0.9386	0.0339	0.8379	0.8979	0.9019
DIS-TE1	0.8925	0.8315	0.8869	0.9186	0.0338	0.8080	0.8713	0.8775
DIS-TE2	0.9163	0.8740	0.9113	0.9411	0.0299	0.8481	0.9046	0.9065
DIS-TE3	0.9269	0.8976	0.9320	0.9565	0.0267	0.8691	0.9235	0.9161
DIS-TE4	0.9096	0.8677	0.9087	0.9440	0.0371	0.8294	0.8993	0.8827
Average	0.9112	0.8673	0.9087	0.9398	0.0323	0.8385	0.8993	0.8969

B FSB dataset documentation

Table 7 summarizes the 60-image benchmark and the agreement between two independent annotations of five representative images.

Table 7: FSB dataset statistics and inter-annotator agreement. Variability is reported as population mean ± standard deviation. The central-corridor share is computed per image as the number of ground-truth foreground pixels inside the middle 50% of the image width divided by the total number of ground-truth foreground pixels. Boundary F-measure uses a three-pixel tolerance.

Dataset property	Value
Images / resolution	60 / 1920×1080
Foreground area ratio	$9.30\% \pm 4.83\%$
Connected components per image	589.4 ± 551.3
Components ≥ 20 px per image	51.3 ± 46.1
Estimated branch width	9.13 ± 5.61 px
Central-corridor share of GT foreground	$81.59\% \pm 23.30\%$
Masks touching an image boundary	39 / 60
Independently relabeled images	5
Minimum IoU / Dice	0.961 / 0.980
Minimum clDice / boundary F-measure	0.951 / 0.945

FSB images were acquired across different locations and capture batches. Table 2 compares methods on all 60 images without target-domain fine-tuning or test-time augmentation.