RETHINKING LANDSLIDE MAPPING AS COARSE PREDICTION UNDER WEAKLY DEFINED SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent deep learning work usually treats landslide mapping as full-resolution semantic segmentation and seeks better performance through stronger backbones, multimodal fusion, or boundary refinement. Yet landslide supervision is often weakly defined: masks can be incomplete, visually interpreted, boundary-ambiguous, and misaligned with the region-level decisions that matter in practice. We revisit the task from a simpler angle: change only the output resolution. Without adding a new architecture, decoder, or inference stage, we vary the spatial granularity of labels and predictions and study how this alone changes performance. Across three datasets, moderate coarse prediction can improve region-level detection, although the gain is dataset-dependent and not monotonic with resolution. To separate useful coarse prediction from trivial task easing, we evaluate direct and pooled fine-resolution predictions on common grids, alongside label-faithfulness and anti-triviality diagnostics. The results point to a dataset-dependent reliable resolution sweet spot and suggest that output granularity is a practical design variable for landslide mapping rather than a minor implementation choice.

1 INTRODUCTION

Landslides are among the most destructive geohazards, and rapid mapping from remote sensing imagery is important for emergency response, hazard assessment, and inventory generation (Jiang et al., 2026; Soares et al., 2020; Xu et al., 2024). Recent progress in deep learning has made semantic segmentation a dominant paradigm for landslide mapping, especially when high-resolution imagery, multi-sensor inputs, or bi-temporal observations are available (Zhang & Wang, 2024; Ghorbanzadeh et al., 2022; Xu et al., 2024; Zhang et al., 2023).

However, pixel-wise landslide segmentation is often more weakly defined than generic semantic segmentation benchmarks. Unlike objects with sharp and stable contours, landslide masks are frequently derived from visual interpretation, incomplete inventories, or post-hoc expert delineation, and can therefore be uncertain, incomplete, or spatially inconsistent (Soares et al., 2020; Böhm et al., 2022; Xiao & Zhang, 2023; Hohn et al., 2025). This mismatch is especially problematic when learning is driven by exact pixel supervision while the underlying evidence is low-resolution, temporally sparse, or only weakly aligned with visually interpretable boundaries.

At the same time, the visual distinction between landslides and background is often subtle. Bare soil, exposed rock, roads, and other disturbed surfaces can share similar spectral or textural patterns with landslide scars, making false positives a recurring error mode even for strong models (Garcia et al., 2023; Tuo et al., 2025; Gao et al., 2026). Small targets, coarse sensor resolution, and boundary ambiguity further amplify the difficulty of enforcing precise pixel-level agreement (Tuo et al., 2025; Wang et al., 2024; Islam et al., 2026). As a result, a model may achieve visually detailed predictions while still overfitting noisy boundaries or unstable supervision, even though many practical applications care more about accurate regional decisions than exact contour fidelity.

These observations motivate us: should landslide mapping always be formulated as full-resolution pixel segmentation? In many practical settings, the target decision is not the exact trace of every boundary pixel, but whether a local region reliably contains landslide evidence. If the supervision is weakly defined and the application is region-oriented, forcing models to predict at the finest possible output resolution may be unnecessarily strict and, in some cases, counterproductive.

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.

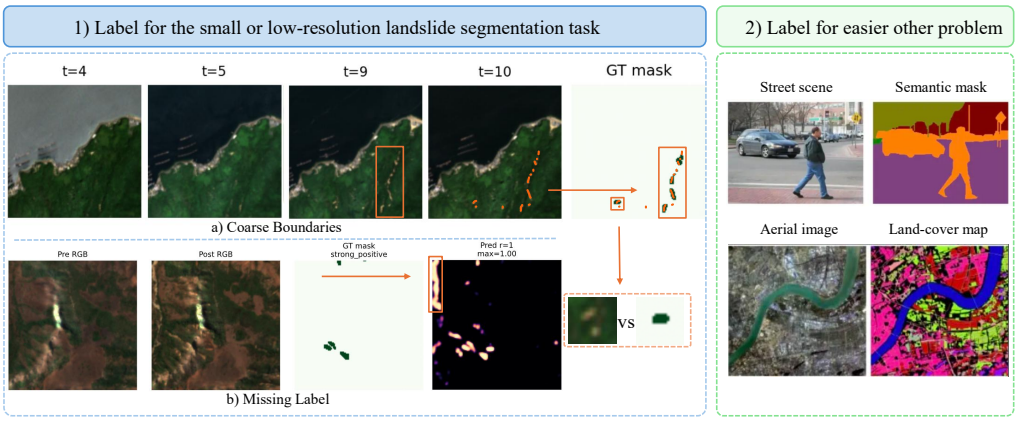


Figure 1: Weak supervision in landslide mapping: boundaries are coarse and some small landslides are unlabeled, creating a mismatch between pixel masks and region-level evidence.

In this work, we study output resolution as a task-definition variable for landslide mapping. We keep the model family largely unchanged and vary only the spatial granularity of the output labels and predictions. This lets us ask whether a very small implementation change can better match weak supervision, reduce brittle boundary fitting, and improve the region-level decisions that downstream use actually depends on.

A second challenge then immediately arises: raw coarse-cell scores alone can be misleading. Extremely coarse outputs may obtain high cell-level F1 or IoU simply because the task becomes easier. We therefore compare direct coarse training with pooled fine predictions on the same evaluation grid and retain label-faithfulness and anti-triviality diagnostics for the spatial trade-off.

Our experiments on three benchmarks show that coarse prediction is indeed useful, but not in the simplistic sense that "coarser is always better". Instead, landslide mapping exhibits a dataset-dependent reliable resolution sweet spot. On weakly defined settings, moderate coarse prediction can improve region-level quality with almost no method overhead; on cleaner or change-driven settings, the best resolution shifts toward finer scales. This reframes output resolution from an implementation detail into a practical design variable for landslide mapping.

Our contributions are summarized as follows:

- We argue that pixel-level landslide segmentation is often weakly defined under imperfect inventories, ambiguous boundaries, and confusing backgrounds, and we formulate output resolution as a task-level design choice rather than a mere implementation detail.
- We show that changing only the output resolution can measurably improve region-level detection quality on three datasets, without introducing a new architecture or extra inference stage.
- We propose a reliability-oriented evaluation protocol for coarse landslide prediction, including indicators for label faithfulness, anti-triviality, and comparison against pooled fine-resolution baselines.
- We show that the favorable resolution range is dataset-dependent, which suggests that output granularity should be explicitly analyzed rather than fixed *a priori*.

2 Related Work

2.1 Landslide Mapping Paradigms

Landslide mapping has long been pursued through visual interpretation, object-oriented analysis, and change detection, and more recently through deep learning-based classification, detection, and segmentation pipelines (Yang et al., 2024; Zhang & Wang, 2024; Cerrudo et al., 2025). Recent reviews

organize landslide deep learning methods into several major task families, including classification, detection, segmentation, sequence modeling, and hybrid frameworks (Zhang & Wang, 2024; Jiang et al., 2026). Among these, semantic segmentation has become especially popular because it offers dense spatial outputs that are directly useful for inventory generation and damage assessment.

2.2 Deep segmentation models for landslide extraction

Recent deep learning studies typically formulate landslide mapping as semantic segmentation, enhancing performance via advanced encoders, multimodal fusion, feature pyramids, attention mechanisms, and decoder refinements (Ghorbanzadeh et al., 2022; Tuo et al., 2025; Wang et al., 2024; Yu et al., 2024; Islam et al., 2026; Hou et al., 2024), with a focus on multi-scale context modeling, boundary refinement, and false alarm suppression (Yu et al., 2024; Tuo et al., 2025; Wang et al., 2024), while all follow the standard task definition of predicting the finest-grained mask. The broader segmentation community has addressed noisy, weakly aligned dense supervision using boundary-aware learning, contour refinement, and robust training with imperfect pseudo-labels (Acuna et al., 2019; Li et al., 2022; Rong et al., 2023); our work complements these approaches by investigating a simpler alternative: coarsening the output space itself for region-level decisions with weakly defined labels, rather than improving the robustness of fine-grained segmentation.

2.3 Datasets, supervision quality, and ambiguity

A parallel thread of work has focused on building larger and more diverse landslide benchmarks, such as Landslide4Sense, CAS Landslide Dataset, Sen12Landslides, GDCLD, DMLD, and GVLM (Ghorbanzadeh et al., 2022; Xu et al., 2024; Hohn et al., 2025; Fang et al., 2024; Chen et al., 2024; Zhang et al., 2023). These datasets have significantly advanced empirical evaluation, but they also make a core difficulty more visible: high-quality landslide masks are expensive to produce, often depend on expert interpretation, and may remain incomplete or noisy even in curated resources (Soares et al., 2020; Böhm et al., 2022; Xiao & Zhang, 2023; Hohn et al., 2025). Notably, the Sen12Landslides data descriptor reports that manual label refinement yields more accurate and less noisy supervision (Hohn et al., 2025). Several studies likewise report that landslide extraction remains difficult because the target is sparse, backgrounds are confusing, and the distinction between landslide and non-landslide surfaces is subtle (Garcia et al., 2023; Tuo et al., 2025; Gao et al., 2026; Zhao et al., 2025). Prior work often observes much lower precision than recall and attributes the gap to false positives from bare soil, roads, exposed rocks, and other visually similar surfaces (Garcia et al., 2023; Tuo et al., 2025; Gao et al., 2026). Together these observations suggest that part of the difficulty may come from asking for overly precise dense decisions under weakly defined supervision.

2.4 Our position

Boundary-label relaxation addresses uncertain pixel annotations (Zhu et al., 2019); our focus is the output decision unit and the resulting trade-off between regional detection and spatial fidelity. Our work is closest to landslide segmentation studies, but differs in focus. We do not propose a new backbone, fusion block, or decoder. Instead, we ask whether output granularity itself should be treated as a design variable when supervision is weakly defined, and whether better coarse results reflect genuinely more reliable predictions rather than mere task easing.

3 Method

3.1 Problem reformulation by output resolution

We consider landslide mapping as binary dense prediction over an input patch $x \in \mathbb{R}^{T\times C\times H\times W}$ with fine mask $m \in \{0,1\}^{H\times W}$. Instead of always optimizing on the native pixel grid, we introduce a task resolution parameter r, where $r = 1$ is the standard pixel-wise setting and $r > 1$ means that one prediction cell covers roughly an $r \times r$ image region. In the implementation, the corresponding output grid size is

$$g_{r,H} = \max\left(1, \operatorname{round}\left(\frac{H}{r}\right)\right), \qquad g_{r,W} = \max\left(1, \operatorname{round}\left(\frac{W}{r}\right)\right), \tag{1}$$

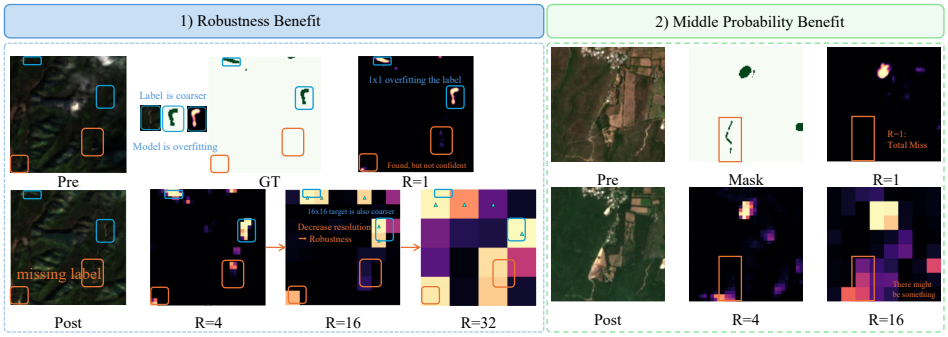


Figure 2: Coarse prediction can be more robust to missing labels and can preserve useful mid-confidence evidence that disappears at 1×1.

so the model predicts a $g_{r,H} \times g_{r,W}$ map rather than an $H \times W$ map.

The key design choice in our code is that this reformulation does not require redesigning the backbone. Any model that outputs a dense logit map $z \in \mathbb{R}^{H \times W}$ can be converted to resolution r by projecting the logits and targets onto the coarse grid. For direct coarse training, the projected logits are

$$\tilde{z}^{(r)} = \begin{cases} z, & r = 1, \\ \mathrm{AdaptiveMaxPool}(z; g_{r,H}, g_{r,W}), & r > 1, \end{cases} \tag{2}$$

where the pooling is adaptive so the same code path also handles settings such as 24×24 and 40×40.

Minimal adaptation. The practical overhead is intentionally tiny. For a base model that already returns a logit map, the resolution reformulation reduces to a few lines:

```
logits = model(x)["segmentation"]
logits_r = adaptive_max_pool2d(logits, (g_h, g_w))
target_r = build_resolution_targets(mask, r)
loss = BCE_Dice(logits_r, target_r)
```

This is one of the main reasons we treat resolution as a task-definition variable: the change is simple enough to be attached to many existing models, datasets, and training pipelines without new architecture engineering.

3.2 Target construction, training objective, and evaluation

For a coarse cell $\Omega_{ij}^{(r)}$, we define the *hard* coarse target by max projection,

$$y_{ij}^{\mathrm{hard}}(r) = \max_{(u,v) \in \Omega_{ij}^{(r)}} m_{uv}, \tag{3}$$

which matches the question "does this region contain landslide evidence?" This is the default training target in the main experiments and also the target used in the direct-versus-pooled fairness comparison.

We additionally compute the *occupancy* target,

$$y_{ij}^{\mathrm{occ}}(r) = \frac{1}{|\Omega_{ij}^{(r)}|} \sum_{(u,v) \in \Omega_{ij}^{(r)}} m_{uv}, \tag{4}$$

which is useful for analyzing how much fine positive area each coarse cell actually contains. We also tested occupancy-based stepped labels and boundary-loss variants as alternatives to hard max labels, but they underperformed the hard target in our setting; details are in Appendix A.2.

The main single-resolution experiments optimize a BCE-Dice objective,

$$\mathcal{L} = (1 - \lambda)\mathcal{L}_{\mathrm{BCE}} + \lambda \mathcal{L}_{\mathrm{Dice}}, \tag{5}$$

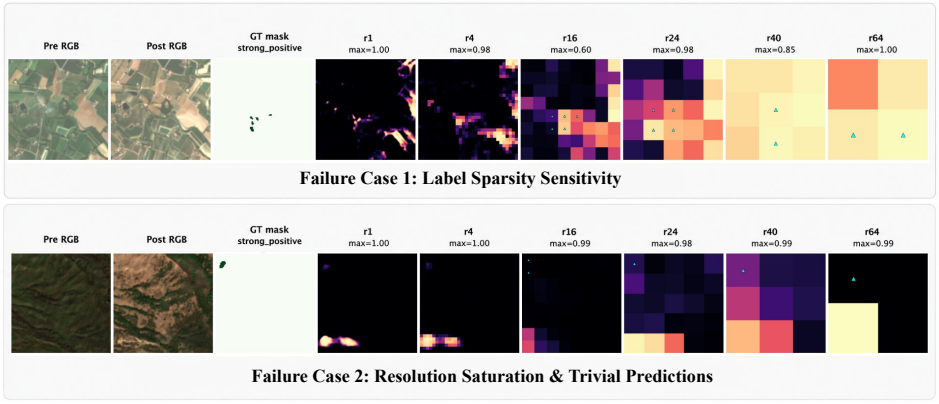


Figure 3: Raw coarse scores can mislead: extreme coarsening may fit missing labels, but can also collapse into trivial block maps.

where $\lambda = 0.5$ in the reported runs. We use Dice smoothing $s = 1.0$ and fixed prediction threshold 0.5; full loss and ablation details are in Appendix A.3.

A central difficulty in coarse prediction is that raw cell-level metrics can improve simply because the task becomes easier. We therefore complement standard cell F1 and cell IoU with three groups of diagnostics. All metrics are computed directly from the evaluation code used in our training scripts.

Label faithfulness. We measure how well coarse positives remain connected to meaningful landslide occupancy using the Brier score and three occupancy-based faithfulness indicators: positive purity, weak positive fraction, and mixed-cell ratio. Together these indicators show whether a resolution creates many coarse positives that are only weakly supported by the underlying fine annotation. Full definitions are given in Appendix A.5.

Anti-triviality. We evaluate whether coarse predictions retain non-trivial spatial structure. Let $\hat{y}_{ij} = \mathbf{1}[p_{ij} \geq 0.5]$. The non-constant prediction rate (NCPR) is the fraction of images whose predicted grid contains at least one positive and at least one negative cell. We also report skill scores against the trivial always-positive predictor at the same grid:

$$\mathrm{IoUS} = \frac{\mathrm{IoU} - \mathrm{IoU}_{\text{all-1}}}{1 - \mathrm{IoU}_{\text{all-1}}}, \qquad \mathrm{F1S} = \frac{\mathrm{F1} - \mathrm{F1}_{\text{all-1}}}{1 - \mathrm{F1}_{\text{all-1}}}. \tag{6}$$

These metrics help identify cases in which extremely coarse outputs obtain high raw scores but collapse into degenerate maps.

Fairness against pooled R1 predictions. Finally, we compare direct coarse training to post-hoc max pooling of the best pixel-wise model, which matches the hard coarse label definition. Coarse max evaluation can raise F1/IoU because fragmented false positives may merge into fewer cells while regional hits are preserved, but this is not a guarantee of a better model; it is exactly why we report pooled baselines and anti-triviality checks. A formal statement is given in Appendix A.1.

4 Experimental Setup

4.1 Datasets

We evaluate on three datasets: Sen12Landslides (Hohn et al., 2025), GVLM (Zhang et al., 2023), and Landslide4Sense (Ghorbanzadeh et al., 2022). Sen12Landslides remains the main development benchmark because its data descriptor explicitly highlights the value of label refinement, while GVLM and Landslide4Sense test whether the same resolution story transfers to different supervision regimes and sensing conditions.

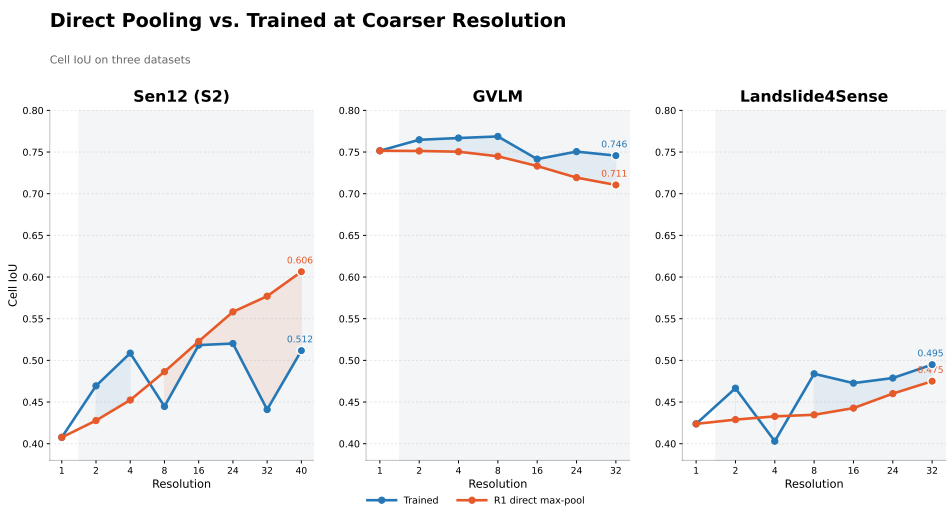


Figure 4: Native-grid resolution sweeps across Sen12, GVLM, and Landslide4Sense. Cell targets change with output resolution, so these curves describe the regional task rather than a fixed-target gain. Table 1 makes the same-target comparison.

Table 1: Direct coarse training and pooled dense predictions evaluated against the same target in each selected comparison. Strict IoU uses the original pixel masks after reconstruction.

Dataset	r	Model	AUPRC	FROC	Comp. F1	Strict IoU
Sen12	2	Direct	.676	.344	.557	.424
		Pooled $r = 1$	.623	.271	.491	.386
GVLM	8	Direct	.947	.692	.563	.715
		Pooled $r = 1$	.935	.686	.535	.716
Landslide4Sense	2	Direct	.698	.370	.515	.469
		Pooled $r = 1$	.623	.228	.418	.429

4.2 Models and Protocol

We do not introduce a new backbone or decoder specialized for coarse prediction. The main sweep uses a 3D U-Net-style encoder-decoder with temporal average pooling, and the main-paper cross-model check uses UNet3D and UTAE; additional configuration details are in Appendix B. Unless otherwise stated, we vary only the output resolution while keeping the rest of the pipeline fixed. The main Sen12 runs use AdamW with learning rate 10^{-3}, weight decay 10^{-2}, up to 50 epochs, 'ReduceLROnPlateau', early stopping, a fixed threshold of 0.5 in the native-grid sweep, gradient clipping at 1.0, and random horizontal/vertical flips plus $90°$ rotations. Patch size is 128×128 on Sen12Landslides and Landslide4Sense, and 256×256 on GVLM; batch size is 16 on Sen12Landslides and Landslide4Sense, and 8 on GVLM. Sen12 additionally uses a 1:1 ratio of clean negatives, so a strong spatial model can have NCPR near 0.5 because pure negative images should be constant negative. The controlled comparison in Table 1 uses an identical evaluation target and samples within each selected dataset/grid, chooses decision thresholds on validation data, and reports AUPRC, FROC partial area over 0–2 false components per image, component F1, and strict original-mask IoU. Predictions are reconstructed to the original pixel grid for the last two metrics. The original native-resolution sweeps remain useful for studying the task change; full setup details are in Appendix B.

5 Results

5.1 Qualitative Motivation

Figure 4 previews how output resolution changes the native regional task across datasets; Table 1 provides the common-target comparison. Figure 1 illustrates the label-side motivation, and Figure 2

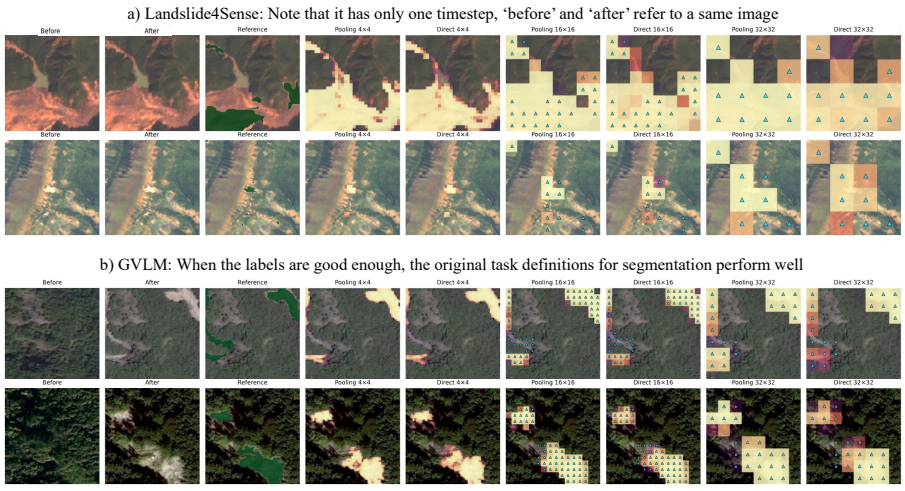


Figure 5: On Landslide4Sense, both pooled 1×1 predictions and directly trained coarse models can produce meaningful region-level outputs when labels are relatively clean.

Table 2: Sen12 reliability analysis. Very coarse outputs score well raw but lose label faithfulness and non-trivial structure. Pur.=positive purity; Weak+=weak positive fraction; Mix=mixed-cell ratio; NCPR=non-constant prediction rate; IoUS/F1S=skill over the trivial all-positive baseline.

Res.	Pur.	Weak+	Mix	NCPR	IoUS	F1S
1×1	1.0000	0.0000	0.0000	0.5070	0.3985	0.5663
4×4	0.5163	0.0905	0.0101	0.4938	0.4495	0.6134
16×16	0.1657	0.5370	0.0448	0.4747	0.4702	0.6193
24×24	0.1044	0.6967	0.0577	0.4883	0.4394	0.5781
32×32	0.0880	0.7445	0.0627	0.4248	0.3257	0.4520
40×40	0.0677	0.8090	0.0678	0.2120	0.3720	0.4921
64×64	0.0498	0.8654	0.0770	0.1585	0.5347	0.6382
128×128	0.0304	0.9243	0.0789	0.0000	0.8499	0.8834

shows that coarser outputs can be more robust to missing labels and preserve useful evidence that is nearly invisible at 1×1. Figure 3 highlights the opposite failure mode: extremely coarse grids may fit missing labels but can also collapse into trivial maps.

5.2 Coarse training under a common evaluation target

Table 1 compares directly trained coarse models with max-pooled $r = 1$ predictions using identical targets and grids. On Sen12 and Landslide4Sense, direct training improves regional AUPRC and component detection while retaining a gain under strict original-mask IoU. GVLM shows a smaller regional gain and essentially unchanged strict pixel IoU. Across 28 model–dataset–resolution comparisons in the extended analysis, direct training improves common-grid IoU in 19 and strict pixel IoU in 17, while reducing false components in 27. These counts also make clear that the benefit is not universal.

The native-resolution sweep in Appendix A documents why this control matters: at larger cells, post-hoc pooling alone can raise F1/IoU without retraining. Our central observation is therefore the regional gain under a shared target, together with the cost in spatial faithfulness as cells grow.

5.3 Reliability analysis reveals a sweet spot

Table 2 explains why the previous result should not be read as "the coarser, the better". Positive purity drops sharply as the grid becomes coarser, while weak positive fraction rises from 0.0905 at 4×4 to 0.9243 at 128×128. In other words, many very coarse positive cells are only weakly supported by

Table 3: GVLM validation. The best resolution shifts toward finer scales. IoUS=IoU skill; Pur.=positive purity; NCPR=non-constant prediction rate.

Res.	F1	IoU	FAR	Miss	IoUS	Pur.	NCPR
1×1	0.8581	0.7515	0.2302	0.0229	0.7339	1.0000	0.4832
4×4	0.8680	0.7668	0.1709	0.0305	0.7488	0.9253	0.4325
16×16	0.8516	0.7415	0.1648	0.0382	0.7161	0.7382	0.4194
24×24	0.8575	0.7505	0.1427	0.0401	0.7227	0.6579	0.4029
32×32	0.8544	0.7457	0.1296	0.0649	0.7142	0.5996	0.3845

Table 4: Landslide4Sense validation. Coarse prediction still improves over 1×1, but the sweet region differs from GVLM and Sen12. IoUS=IoU skill; Pur.=positive purity; NCPR=non-constant prediction rate.

Res.	F1	IoU	FAR	Miss	IoUS	Pur.	NCPR
1×1	0.5953	0.4238	0.6250	0.0541	0.4127	1.0000	0.8400
4×4	0.5745	0.4030	0.5417	0.1213	0.3839	0.6110	0.7675
16×16	0.6420	0.4727	0.6061	0.0317	0.4213	0.2127	0.8488
24×24	0.6476	0.4788	0.2311	0.2276	0.3889	0.1282	0.5925
32×32	0.6621	0.4949	0.6061	0.0578	0.3818	0.1033	0.8288

the underlying fine annotation. At 128×128, NCPR is structurally forced to 0 because the 128×128 patch becomes a single prediction cell; together with the purity drop, this marks the extreme setting as an image-level presence task rather than a useful spatial map.

These native-grid patterns describe the regional task at different cell sizes. Moderate resolutions such as 16×16 and 24×24 improve raw performance while maintaining more non-trivial structure than extreme coarsening; the same-target comparison and validation selector can favor finer operating points. Very coarse resolutions can look strong under raw cell metrics, but the accompanying degeneration in purity and non-constant structure indicates that they are not the most reliable operating point.

5.4 Hard labels remain the main answer

An intuitive alternative is to replace hard max labels with occupancy-based stepped labels or boundary-aware losses. The tested variants do not improve this setting. This does not exclude other robust losses, but it shows that the gains here do not require the particular soft-label variants we evaluated. Detailed formulations and ablations are provided in Appendices A.2 and C.

5.5 External datasets show a dataset-dependent sweet spot

Figure 5 and Tables 3–4 show that the favorable resolution range depends on the dataset. The provided qualitative examples come from Landslide4Sense and compare direct coarse training against pooled 1×1 predictions under several resolutions. They suggest that when labels are good enough, multiple segmentation-style task definitions can all work reasonably well, but the exact balance between pooling and direct coarse training still depends on data properties and scale.

The combined message is important: our paper is not claiming a universal optimum such as 16×16 or 24×24. Instead, we argue that landslide mapping has a *reliable sweet region* whose location depends on data properties such as label quality, target scale, and background ambiguity.

5.6 Additional analyses

The controlled extension is not confined to the original U-Net family: ResUNet on Landslide4Sense and SegFormer on GVLM improve common-grid IoU over their own $r = 1$ settings by .0228 and .0296 at validation-selected resolutions. A two-seed Temporal SegFormer comparison on Sen12 yields a mean common-grid IoU gain of .0267. Replications across all settings are mixed, so these

results establish breadth of the observed effect rather than architecture-independent dominance. Supplementary Table S6 retains the earlier cross-model and modality sweep.

Controlled label perturbations. To test the weak-supervision motivation more directly, we use two existing controls that perturb training annotations while keeping validation and test targets unchanged. On GVLM with training masks shifted by eight pixels, direct training at $r = 8$ reaches FROC 0.774, component F1 0.656, and 0.413 false components per image, versus 0.755, 0.624, and 0.528 for shifted dense training followed by pooling. Boundary relaxation preserves strict pixel fidelity better in this setting. In a separate false-positive-blob pilot, direct $r = 2$ training exceeds pooled $r = 1$ by 0.026 common-grid IoU and 0.038 AUPRC, with 0.216 fewer false components per image. These controls support a regional-screening advantage under the tested annotation errors, while also showing the cost of giving up precise boundaries. Details are in Appendix F.

Pooling and label-target controls. The full comparisons in Appendix E separate gains from changing the region-level decision from gains due to training a model at coarse resolution. On Sen12, post-hoc max pooling of the 1×1 prediction recovers much of the apparent benefit and can exceed direct coarse training at several resolutions. Average pooling instead deteriorates as cells grow because sparse landslide evidence is diluted before thresholding. The other datasets show different regimes: GVLM gives direct coarse training only a small local advantage near the native grid, while on Landslide4Sense direct training becomes more useful from 16×16 onward. Thus a coarse decision unit is a substantial part of the gain, and retraining should be credited separately when it improves beyond the pooled fine model.

Appendix C provides a complementary target check. On Sen12, occupancy-based stepped soft labels underperform the hard max target at every tested resolution, especially at intermediate and coarse grids. This result does not make hard targets universally preferable, but it shows that simply softening weakly occupied cells does not recover the missed positive evidence in this setting.

6 Discussion & Limitations

Our results do not claim pixel-wise segmentation is invalid for landslide mapping. Rather, it can be over-constrained relative to supervision quality and application needs, while moderately coarse predictions may better match practitioners' regional decision-making. Reliability analysis also shows why "coarser is not always better": very coarse outputs score well on raw cell metrics but become less faithful to underlying occupancy and less spatially informative. This work is limited in that it does not create new visual detail or recover truly better boundaries, serving only as a practical surrogate for weak labels rather than a full solution to annotation quality. Our validation-only screening rule selects among candidate resolutions using common-grid AUPRC, excludes grids smaller than 8×8 and NCPR below 0.30, and breaks near ties toward finer outputs. It is intended for regional screening; boundary-sensitive use still requires strict pixel and component checks. Future work should connect the selected scale more directly to target geometry and annotation uncertainty.

7 Conclusion

We studied landslide mapping through the lens of output resolution. With the model family largely fixed and only prediction granularity varied, we showed that simple task reformulation improves region-level detection under weakly defined supervision, while overly coarse outputs risk degenerating into trivial block maps. A reliability-oriented evaluation further revealed the optimal scale is dataset-dependent rather than universal. Overall, output granularity should be treated as a practical task-design variable in landslide mapping, especially when fine boundaries are noisy, incomplete, or weakly supported by data.

AI USE STATEMENT

Generative AI tools were used to aid and polish writing, retrieve and discover related work, assist with research ideation and execution, and draft sections of this paper. They also assisted with adapting the manuscript to the ICLR 2027 template and checking figure and table placement and PDF layout. The authors take responsibility for the final text, citations, experiments, and claims.

REFERENCES

David Acuna, Amlan Kar, and Sanja Fidler. Devil is in the edges: Learning semantic boundaries from noisy annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11075–11083, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Acuna_Devil_Is_in_the_Edges_Learning_Semantic_Boundaries_From_Noisy_CVPR_2019_paper.html.

Vanessa Böhm, Wei Ji Leong, Ragini Bal Mahesh, Ioannis Prapas, Edoardo Nemni, Freddie Kalaitzis, Siddha Ganju, and Raul Ramos-Pollan. SAR-based landslide classification pretraining leads to better segmentation, 2022. URL https://arxiv.org/abs/2211.09927.

Samuel Cerrudo, Daniel De Castro, and Prospero C. Jr. Naval. Automated landslide detection: A comparative study of change detection and semantic segmentation techniques. In *Neural Information Processing*, volume 15295 of *Lecture Notes in Computer Science*, pp. 180–194. Springer, 2025. URL https://link.springer.com/book/10.1007/978-981-96-6603-4.

Jie Chen, Xu Zeng, Jingru Zhu, Ya Guo, Liang Hong, Min Deng, and Kaiqi Chen. The diverse mountainous landslide dataset (DMLD): A high-resolution remote sensing landslide dataset in diverse mountainous regions. *Remote Sensing*, 16(11):1886, 2024. doi: 10.3390/rs16111886. URL https://www.mdpi.com/2072-4292/16/11/1886.

Chengyong Fang, Xuanmei Fan, Xin Wang, Lorenzo Nava, Hao Zhong, Xiujun Dong, Jixiao Qi, and Filippo Catani. A globally distributed dataset of coseismic landslide mapping via multi-source high-resolution remote sensing images. *Earth System Science Data*, 16(10):4817–4842, 2024. doi: 10.5194/essd-16-4817-2024. URL https://essd.copernicus.org/articles/16/4817/2024/essd-16-4817-2024.html.

Mengjie Gao, Fang Chen, Lei Wang, and Bo Yu. A spike-inspired adaptive spatial suppression framework for large-scale landslide extraction. *Remote Sensing*, 18(1):129, 2026. doi: 10.3390/rs18010129. URL https://www.mdpi.com/2072-4292/18/1/129.

Guilherme P. B. Garcia, Lucas P. Soares, Mateus Espadoto, and Carlos H. Grohmann. Relict landslide detection using deep-learning architectures for image segmentation in rainforest areas: a new framework. *International Journal of Remote Sensing*, 44(7):2168–2195, 2023. doi: 10.1080/01431161.2023.2197130. URL https://www.tandfonline.com/doi/full/10.1080/01431161.2023.2197130.

Omid Ghorbanzadeh, Yonghao Xu, Pedram Ghamisi, Michael Kopp, and David P. Kreil. Landslide4sense: Reference benchmark data and deep learning models for landslide detection. *IEEE Transactions on Geoscience and Remote Sensing*, 60:1–17, 2022. doi: 10.1109/TGRS.2022.3215209. URL https://doi.org/10.1109/TGRS.2022.3215209.

Paul Hohn, Konrad Heidler, Robert Behling, and Xiao Xiang Zhu. A spatio-temporal dataset for satellite-based landslide detection. *Scientific Data*, 12(1):1772, 2025. doi: 10.1038/s41597-025-06167-2. URL https://doi.org/10.1038/s41597-025-06167-2.

Changhong Hou, Junchuan Yu, Daqing Ge, Liu Yang, Laidian Xi, Yunxuan Pang, and Yi Wen. TransLandSeg: A transfer learning approach for landslide semantic segmentation based on vision foundation model, 2024. URL https://arxiv.org/abs/2403.10127.

Md Minhazul Islam, Yunfei Yin, Md Tanvir Islam, Zheng Yuan, and Argho Dey. A dual-stream framework for landslide segmentation with cross-attention enhancement and gated multimodal fusion. *Computers, Materials & Continua*, 86(3):8, 2026. doi: 10.32604/cmc.2025.072550. URL https://www.techscience.com/cmc/v86n3/65473/html.

Pan Jiang, Zhengjing Ma, and Gang Mei. Review article: Deep learning for potential landslide identification: data, models, applications, challenges, and opportunities. *Natural Hazards and Earth System Sciences*, 26:487–529, 2026. doi: 10.5194/nhess-26-487-2026. URL https://nhess.copernicus.org/articles/26/487/2026/nhess-26-487-2026.html.

Jing Li, Junsong Fan, and Zhaoxiang Zhang. Towards noiseless object contours for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16856–16865, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Li_Towards_Noiseless_Object_Contours_for_Weakly_Supervised_Semantic_Segmentation_CVPR_2022_paper.html.

Shenghai Rong, Bohai Tu, Zilei Wang, and Junjie Li. Boundary-enhanced co-training for weakly supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19574–19584, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Rong_Boundary-Enhanced_Co-Training_for_Weakly_Supervised_Semantic_Segmentation_CVPR_2023_paper.html.

Lucas P. Soares, Helen C. Dias, and Carlos H. Grohmann. Landslide segmentation with U-Net: Evaluating different sampling methods and patch sizes, 2020. URL https://arxiv.org/abs/2007.06672.

Wanbing Tuo, Jin Zeng, Fengmin Wu, Wei Wu, Xinchao Wang, and Xingchang Zhao. Landslide detection using multimodal data fusion and an improved DeepLabv3+ model. *Scientific Reports*, 16(1):1383, 2025. doi: 10.1038/s41598-025-31208-6. URL https://doi.org/10.1038/s41598-025-31208-6.

Kai Wang, Daojie He, Qingqiang Sun, Lizhi Yi, Xiaofeng Yuan, and Yalin Wang. A novel network for semantic segmentation of landslide areas in remote sensing images with multi-branch and multi-scale fusion. *Applied Soft Computing*, 158:111542, 2024. doi: 10.1016/j.asoc.2024.111542. URL https://www.sciencedirect.com/science/article/pii/S1568494624003168.

Te Xiao and Li-Min Zhang. Data-driven landslide forecasting: Methods, data completeness, and real-time warning. *Engineering Geology*, 317:107068, 2023. ISSN 0013-7952. doi: 10.1016/j.enggeo.2023.107068. URL https://www.sciencedirect.com/science/article/pii/S0013795223000856.

Yulin Xu, Chaojun Ouyang, Qingsong Xu, Dongpo Wang, Bo Zhao, and Yutao Luo. CAS landslide dataset: A large-scale and multisensor dataset for deep learning-based landslide detection. *Scientific Data*, 11:12, 2024. doi: 10.1038/s41597-023-02847-z. URL https://doi.org/10.1038/s41597-023-02847-z.

Zhiqiang Yang, Wenwen Qi, Chong Xu, and Xiaoyi Shao. Exploring deep learning for landslide mapping: A comprehensive review. *China Geology*, 7(2):330–350, 2024. doi: 10.31035/cg2024032. URL https://www.sciencedirect.com/science/article/pii/S2096519224001137.

Chih-Chang Yu, Yuan-Di Chen, Hsu-Yung Cheng, and Chi-Lun Jiang. Semantic segmentation of satellite images for landslide detection using foreground-aware and multi-scale convolutional attention mechanism. *Sensors*, 24(20):6539, 2024. doi: 10.3390/s24206539. URL https://www.mdpi.com/1424-8220/24/20/6539.

Qi Zhang and Teng Wang. Deep learning for exploring landslides with remote sensing and geo-environmental data: Frameworks, progress, challenges, and opportunities. *Remote Sensing*, 16(8):1344, 2024. doi: 10.3390/rs16081344. URL https://www.mdpi.com/2072-4292/16/8/1344.

Xiaokang Zhang, Weikang Yu, Man-On Pun, and Wenzhong Shi. Cross-domain landslide mapping from large-scale remote sensing images using prototype-guided domain-aware progressive representation learning. *ISPRS Journal of Photogrammetry and Remote Sensing*, 197:1–17, 2023. doi: 10.1016/j.isprsjprs.2023.01.018. URL https://www.sciencedirect.com/science/article/pii/S0924271623000242.

Chengwei Zhao, Long Li, Yubo Wang, Xuqing Li, Chong Xu, Yubin Song, Dongsheng Ren, and Cheng Xiao. D2FLS-Net: Dual-stage DEM-guided fusion transformer for landslide segmentation. *PLOS ONE*, 20(11):e0337412, 2025. doi: 10.1371/journal.pone.0337412. URL https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0337412.

Yi Zhu, Karan Sapra, Fitsum A. Reda, Kevin J. Shih, Shawn Newsam, Andrew Tao, and Bryan Catanzaro. Improving semantic segmentation via video propagation and label relaxation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8856–8865, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Zhu_Improving_Semantic_Segmentation_via_Video_Propagation_and_Label_Relaxation_CVPR_2019_paper.html.