

Article

GeoBASP: Geography-Aware Structured Pruning for Landslide-Scar Segmentation in the Nepal Himalaya

Deepak Ghimire 

AI Application Research Center, Korea Electronics Technology Institute, Jeonju 54853, Republic of Korea; deepak@keti.re.kr

Abstract

Rapid landslide mapping is important for disaster response, yet modern segmentation networks can require substantial computation and storage. This study proposes Geography-, Boundary-, and Scar-Aware Structured Pruning (GeoBASP) for compact landslide-scar segmentation. GeoBASP measures the importance of removable channel groups from four complementary aspects: complete-mask accuracy, boundary delineation, small-scar detection, and the rejection of difficult non-landslide areas. These signals are combined across training blocks within the Koshi Basin using robust statistics to guide pruning under measured computational budgets. Experiments used a spatially separated dataset derived from the ICIMOD Koshi Basin 2024 inventory, Sentinel-2 imagery, and terrain slope. ResNet-50 U-Net, ConvNeXt-V2-Tiny U-Net, and SegFormer-B2 were pruned at 25%, 50%, and 70% multiply-accumulate (MAC) reduction. Two general-purpose baselines were evaluated across all three architectures and one segmentation-oriented method on the two U-Nets. GeoBASP achieved the highest overall mean validation Dice in the three-architecture comparison and exceeded the segmentation-oriented method at 25% and 50% reduction on both U-Nets. Rankings varied with architecture and compression level. Only GeoBASP satisfied the predefined non-inferiority condition, with qualifying runs restricted to the 25% target; no method satisfied it at 50% or 70% reduction. For the highest-performing dense architecture, adding slope to the imbalance-aware spectral input increased the mean test Dice from 0.471 ± 0.007 to 0.520 ± 0.004 . At the moderate operating point, GeoBASP reduced MACs by 24.2% and parameters by 12.3% while retaining a test Dice of 0.505 compared with 0.516 for the dense model. For ConvNeXt-V2-Tiny U-Net, validation Dice remained 0.461 ± 0.008 at approximately 50% MAC reduction and 0.419 ± 0.004 at 70%, with corresponding parameter reductions of 33.3% and 53.7%. These findings support moderate compression within the evaluated basin, while performance depends on architecture and aggressive pruning incurs accuracy losses.



Academic Editors: D. J. Lee and Dong Zhang

Received: 21 September 2026

Revised: 14 October 2026

Accepted: 16 October 2026

Published: 17 October 2026

Copyright: © 2026 by the author.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Keywords: landslide segmentation; Sentinel-2; Nepal Himalaya; structured channel pruning; model compression; remote sensing; spatial validation

1. Introduction

Landslides are a serious hazard in the Nepal Himalaya. Steep slopes, monsoon rainfall, earthquakes, road construction, and changing land use create conditions for frequent slope failure. Landslide inventories are needed for emergency response and later hazard assessment, but field mapping and manual image interpretation require considerable time and skilled personnel [1,2]. Satellite remote sensing can support more regular mapping over large and difficult terrain. Sentinel-2 is particularly useful because it provides visible,

near-infrared (NIR), and short-wave infrared (SWIR) measurements with frequent regional coverage [3].

Automatic mapping can reduce this effort, but the problem remains difficult. Landslide scars occupy a small part of most images, vary greatly in size and shape, and can be confused with roads, river banks, bare soil, cloud, and snow. A useful model must detect sparse scars, preserve their boundaries, and avoid false alarms when it is applied away from the training area. Terrain information can support this decision, but an additional input alone does not solve the strong class imbalance.

Accurate segmentation networks also require substantial computation and model storage. This limits repeated large-area mapping and future use on field computers, small servers, or edge devices. General structured-pruning criteria can reduce this cost, but their average importance scores may remove a feature that is uncommon yet necessary for a small scar, an uncertain boundary, or one geographic subregion. The dense-model adaptation and the compression method should therefore reflect the same mapping problem.

This study introduces GeoBASP to address this gap. The main idea is to protect channel groups that are important for at least one landslide-specific case: the complete mask, the scar boundary, small scars, or difficult background. Importance is measured over separate geographic blocks so that a channel is not removed only because it is uncommon in one local area. The resulting scores rank removable dependency groups under measured MAC-reduction budgets, connecting mapping priorities with the objective of a smaller and less computationally demanding model. Here, geography awareness refers to calibration and aggregation across spatial blocks within the Koshi Basin. The experiments assess generalization to held-out basin sectors; transfer to a geographically distinct region is outside the evaluated scope.

The main contributions are: (1) a terrain-aware Sentinel-2 dataset derived from the ICIMOD Koshi Basin 2024 inventory, with spatially separated training, validation, and test regions; (2) a terrain- and imbalance-aware segmentation design evaluated with residual, modern convolutional, and Transformer backbones; (3) GeoBASP, a post-training structured-pruning criterion that combines normalized sensitivity for four landslide-mapping regimes, training-block median-MAD aggregation, and dependency-group ranking; and (4) comparisons at three MAC-reduction targets under shared structural constraints and recovery settings, using group magnitude and group Taylor across all three architectures and the segmentation-oriented SIRFP on the two U-Nets.

2. Related Work

2.1. Landslide Detection and Semantic Segmentation

Early automatic landslide mapping commonly used object-based image analysis and conventional classifiers. For example, Stumpf and Kerle combined image objects with random forests to map landslides from remote-sensing data [4]. Deep learning later improved the ability to learn irregular spatial patterns directly from imagery. Comparative studies have shown that semantic-segmentation networks can outperform traditional machine-learning approaches for landslide mapping [5]. Encoder-decoder methods such as U-Net [6] and DeepLab-style models [7] are particularly suitable because they combine broad context with a dense pixel output.

Research in Nepal has included national-scale detection from Landsat imagery [8] and semantic segmentation of landslide scars using U-Net [9]. These studies demonstrate the potential of deep learning, but performance can depend strongly on image resolution, label quality, and geographic partitioning. Unsupervised Sentinel-2 detection also shows that learning transferable landslide features remains difficult [10]. The present study therefore uses a fixed 20 km spatial separation and reports both pixel overlap and scar-size behaviour.

2.2. Terrain-Aware and Multisource Segmentation

Optical reflectance describes vegetation loss, exposed soil, and surface moisture, while terrain provides information about the physical setting of slope failure. Two-stream CNNs have combined spectral and topographic evidence for landslide mapping [11]. Landslide4Sense further established a multisource benchmark containing Sentinel-2 bands, elevation, and slope and evaluated several segmentation architectures [12]. These results support multisource learning, but they do not establish that every terrain channel improves every architecture or geographic region.

The architectures considered here represent three model families. ResNet provides a conventional residual CNN [13]; ConvNeXt-V2 updates convolutional design and pre-training while retaining a CNN structure [14]; and SegFormer combines a hierarchical Transformer encoder with a compact multiscale decoder [15]. In this study, their standard backbones are adapted to four-channel landslide input and binary output. The adaptation is evaluated separately from pruning so that any compression result is interpreted relative to a suitable dense model.

2.3. Structured Pruning for Dense Prediction

Structured pruning removes complete filters, channels, or dependent parameter groups and produces a smaller dense network that can run on general hardware [16–18]. Magnitude pruning ranks parameters by weight norm, whereas Taylor pruning estimates the first-order change in loss after removal [19]. Reviews show that pruning results depend on the architecture, training procedure, target reduction, and recovery strategy; matched comparisons are therefore important [20,21]. Most early criteria were developed for image classification. Segmentation-oriented methods already incorporate spatial information. CAP calculates channel affinities from feature maps to guide a sparsity-inducing stage on batch-normalization scaling factors; channels are then selected using the resulting factors, and the pruned model is fine-tuned [22]. SIRFP measures pairwise redundancy of unpooled spatial feature maps and formulates retained-channel selection as a maximum-edge-weight clique problem, approximated by greedy channel removal. Its published pipeline accumulates graph weights through an exponential moving average during training and uses progressive pruning and fine-tuning [23].

Other structured pruning studies have investigated mixed magnitude–similarity criteria, automatic criterion and layer selection, and joint training–pruning schedules [24–26]. GeoBASP defines importance through supervised loss sensitivity for four explicit regimes: the complete mask, scar boundaries, small scars, and hard negatives. Unit-gate derivatives are normalized within each layer, aggregated by median plus MAD across training blocks within the same basin, and combined by a maximum across regimes. Scores from coupled output channels are then combined for dependency-group ranking. This design differs from CAP’s contextual guidance for sparsification and SIRFP’s pairwise feature-redundancy objective: its spatial aggregation operates across geographic training blocks, while its task signals measure declared mapping losses. The contribution is the specified combination of regimes and block-level aggregation in a landslide-pruning criterion. First-order sensitivity [19] and dependency tracing and physical removal [27] are established components used by the method.

3. Materials and Methods

3.1. Study Design and Dataset Preparation

3.1.1. Problem Definition

The task is binary semantic segmentation of mapped landslide scars. Given an input tensor $\mathbf{x} \in \mathbb{R}^{4 \times H \times W}$, the network predicts a probability $p_{ij} \in [0, 1]$ that each 10 m pixel

(i, j) belongs to a landslide. This is neither whole-image classification nor future landslide susceptibility prediction. A separate binary supervision mask excludes pixels that cannot be treated reliably as either landslide or non-landslide; excluded pixels do not contribute to the loss or accuracy metrics.

3.1.2. Landslide Inventory and Satellite Imagery

The primary label source is the ICIMOD Landslide Inventory of Koshi Basin 2024 [28]. The dataset contains 9210 valid polygons in WGS 84/UTM zone 45N (EPSG:32645). Each polygon records an associated Sentinel-2 acquisition, enabling image–label alignment. The study uses the corresponding Sentinel-2 Level-2A surface-reflectance products.

The network input comprises B11 (SWIR1), B08 (NIR), B04 (red), and terrain slope derived from the Copernicus DEM GLO-30. B11 was bilinearly resampled from 20 m to the 10 m B04 reference grid; categorical scene classification was resampled by nearest neighbor. Radiometric consistency between the cloud-optimized GeoTIFF product and an original ESA B04 JP2 sample was verified before a reflectance scale of 0.0001, and no additional offset was applied. Normalization statistics were computed from supervised pixels in the training partition only. DEM tiles were reprojected to each Sentinel-2 patch with a two-pixel halo, and slope in degrees was calculated on the aligned 10 m surface. The slope channel was normalized using training-partition statistics only.

3.1.3. Study Area and Spatial Partitioning

Figure 1 shows the study patches and spatial partitions in the transboundary Koshi Basin. Most patches lie in Nepal; the dataset also contains scar-containing and background-only patches in China, and background-only patches in India. These partitions assess within-basin generalization; cross-region transfer remains untested.

Source patches contain 512×512 pixels. Positive patches cover all inventory polygons, and fixed negative-only patches represent confirmed background, including difficult bare-soil and water cases. The combined dataset composition is summarized in Table 1. The partitions are geographically separated by complete 20 km guard columns, following the principle that spatial prediction should be assessed outside the local neighborhood used for model fitting [29]; pairwise polygon overlap is zero, and the minimum patch-footprint separation is 20.48 km (Figure 1). The test partition remains untouched during architecture, threshold, and pruning selection.

Table 1. Composition of the spatially separated patch dataset. Guard patches are excluded from model fitting and evaluation.

Partition	Positive	Negative	Total	Unique Polygons	Spatial Blocks
Training	708	283	991	5050	101
Validation	130	52	182	1338	26
Held-out test	131	52	183	1202	22
Excluded guard	169	0	169	1626	15

3.1.4. Target and Supervision Masks

Inventory polygons were rasterized on the 10 m reference grid. The exact systematic mapping footprint was not supplied with the inventory. The primary supervision domain was therefore reconstructed from valid Sentinel-2 observations within the referenced MGRS tile cores, inside the Koshi Basin boundary, and at elevations below 4500 m; every inventory polygon was retained as a positive override. A second domain without the elevation threshold is retained for sensitivity analysis. This is treated as a study-specific supervision domain rather than an official ICIMOD mapping footprint.

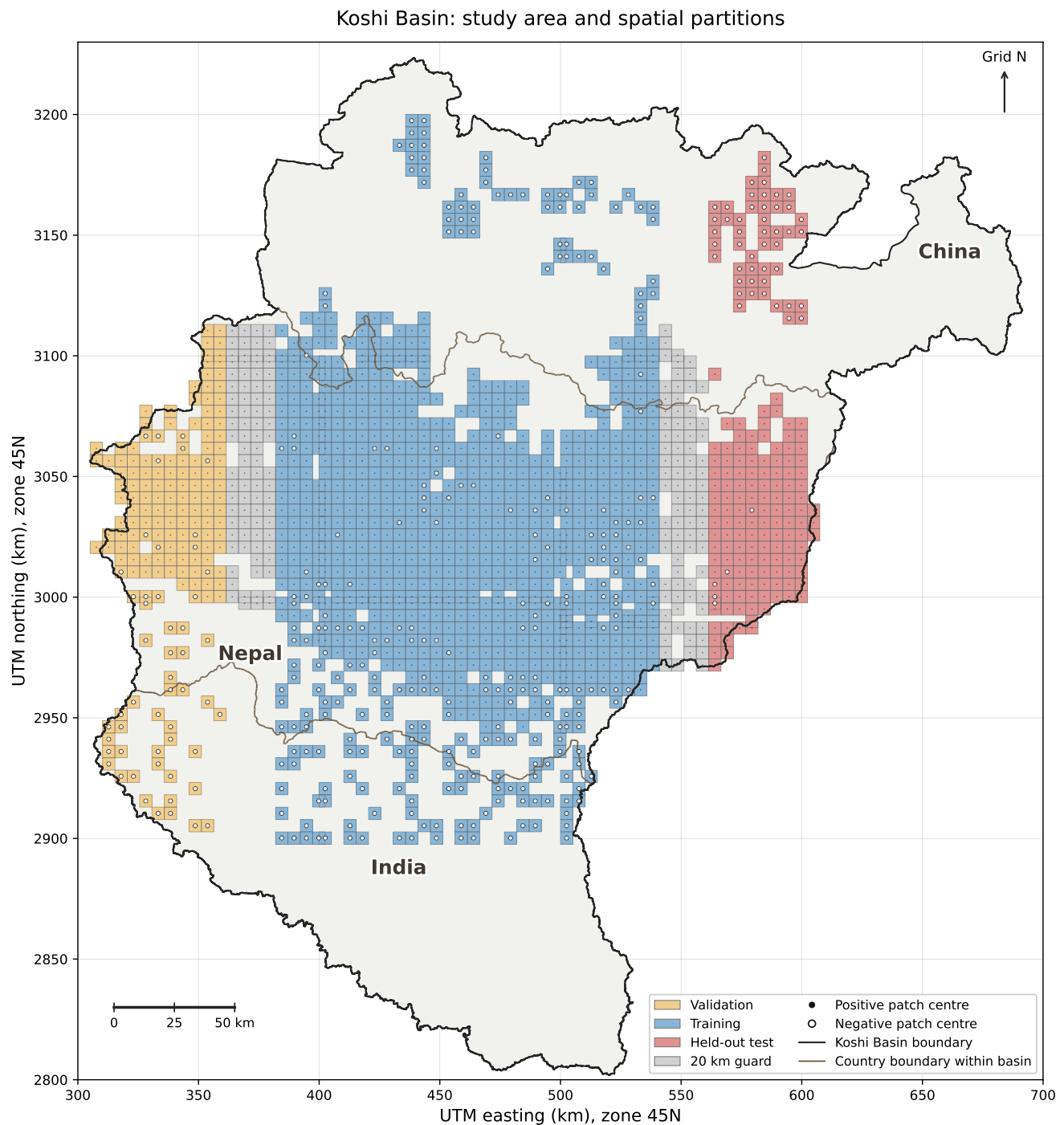


Figure 1. Study area and geographic partitions in the Koshi Basin. The black outline shows the full basin; brown lines and labels identify Nepal, China, and India using ICIMOD boundary data. Colored 5.12×5.12 km patch footprints represent validation, training, held-out test, and excluded 20 km guard regions. Grey outlines delineate individual tiles; filled and open points denote positive-patch and fixed negative-patch centers, respectively. Grid north and the 50 km scale bar use WGS 84/UTM zone 45N.

Pixels classified as no-data, saturated or defective, cast shadow, cloud shadow, medium- or high-probability cloud, cirrus, or snow and ice were excluded. The resulting training target has three operational states: landslide, supervised non-landslide, and ignored. Ignored is not a prediction class. Figure 2 illustrates the source patches, cropped inputs, rasterized ground-truth landslide scars, and three-state supervision masks used for training.

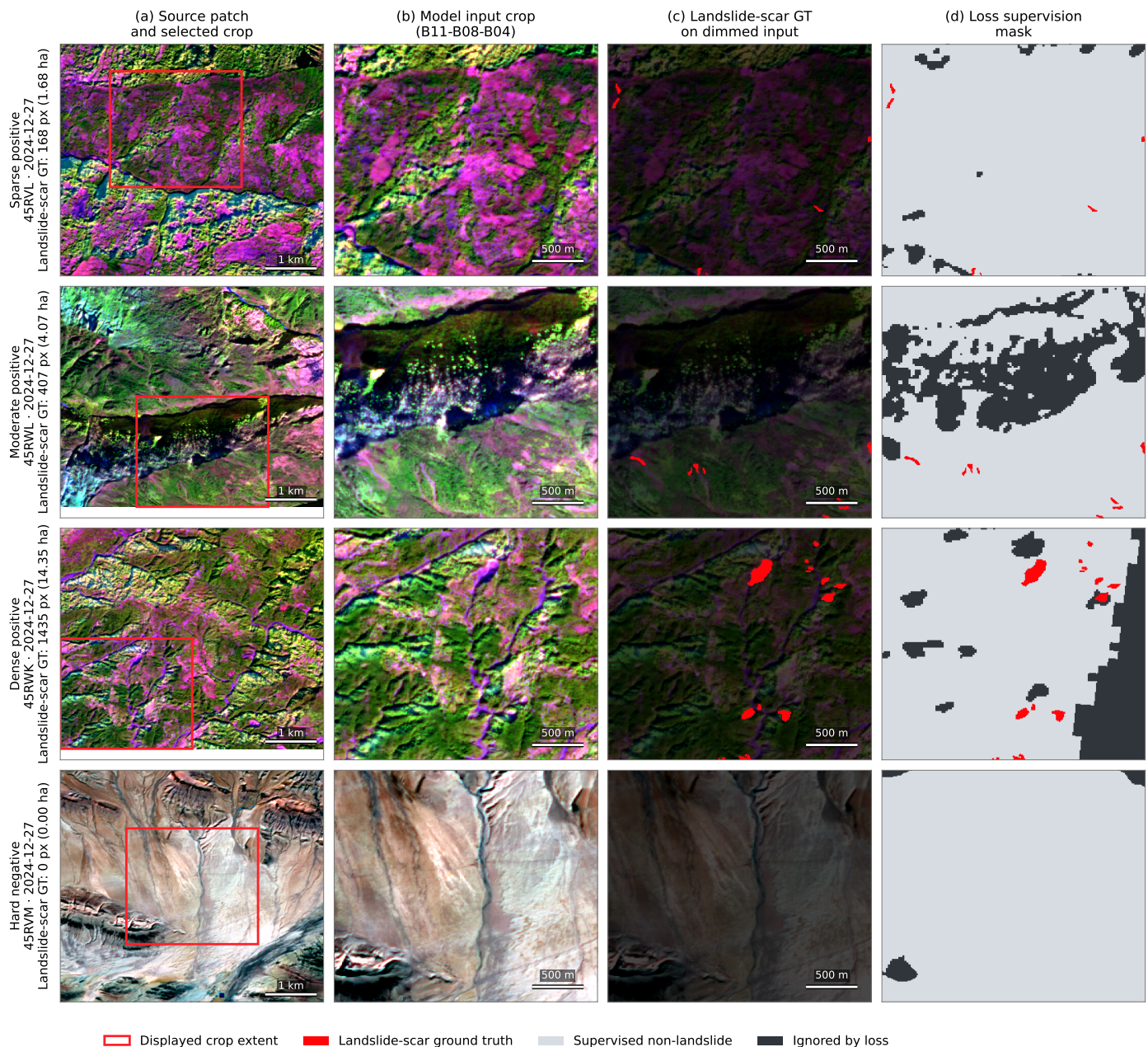


Figure 2. Representative training samples for binary landslide semantic segmentation. (a) Full 512×512 pixel Sentinel-2 source patches (5.12×5.12 km), with the displayed crop outlined in red; (b) enlarged 256×256 pixel views of the outlined areas; (c) rasterized landslide-scar ground-truth pixels in bright red over deliberately dimmed imagery; (d) the three-state supervision supplied to the loss. The input is a contrast-stretched B11–B08–B04 (SWIR–NIR–red) composite at 10 m resolution. The first three rows span sparse to dense positive patches; the fourth shows a bare-ground hard negative. All examples are from the training partition.

During training, 384×384 crops are sampled from source patches. The patch sampler targets 80% positive patches, and a positive crop is centered near a supervised landslide pixel with probability 0.90. Random right-angle rotations and flips provide label-preserving augmentation. Independent per-channel reflectance gains in $[0.95, 1.05]$, offsets in $[-0.005, 0.005]$, and Gaussian noise with standard deviation 0.001 provide mild radiometric augmentation. Validation uses complete 512×512 patches with the natural class distribution.

3.2. Dense Segmentation Model and Training

The dense model family comprises U-Net-style decoders with ImageNet-initialized ResNet-50 [13] and ConvNeXt-V2-Tiny [14] encoders, together with SegFormer-B2 [15] and its multiscale decoder. These architectures cover a conventional residual CNN, a modern convolutional model, and a Transformer-based segmentation model. The U-Net decoders restore spatial detail through skip connections and use 0.10 dropout. The pretrained three-channel input layer is expanded to accept slope as a fourth channel. The new slope weights start at zero, so the spectral representation is preserved at the beginning of training and terrain information is learned gradually.

Class imbalance is addressed using class-balanced focal loss [30] and Tversky loss [31]. Their weights are 0.40 and 0.60, respectively. The foreground focal weight is 0.75, and the Tversky false-negative and false-positive weights are 0.70 and 0.30. These settings give greater attention to missed landslide pixels while keeping false alarms under control. Boundary-focused supervision is also relevant to remote-sensing segmentation because region overlap alone can hide poor scar outlines [32]. Optimization uses AdamW, a batch size of 8, automatic mixed precision, gradient clipping, an eight-epoch warm-up, and cosine learning-rate decay. Encoder and decoder learning rates are 3×10^{-5} and 3×10^{-4} .

An exponential moving average with decay 0.999 provides a stable model for validation. Training continues for at most 250 epochs and stops when validation performance no longer improves. For each architecture, the model with the largest validation Dice is retained. Its final probability threshold is also selected on validation data and is not adjusted using test images. The implementation uses PyTorch [33].

The contribution of the dense-model adaptations was assessed with three ConvNeXt-V2-Tiny U-Net configurations. The standard spectral configuration used B11, B08, and B04 with empirical-distribution sampling and equally weighted binary cross-entropy and Dice losses. The imbalance-aware configuration retained the same bands but adopted foreground-aware patch and crop sampling together with the focal–Tversky objective. The terrain-aware configuration additionally included slope while preserving the imbalance-aware training settings. Probability thresholds were determined independently from the corresponding validation data and applied unchanged to the held-out test region.

3.3. Post-Training Structured Pruning

Pruning is applied after dense-model training. Torch-Pruning and DepGraph identify dependency groups whose channels must be removed jointly [27]. Channels are removed physically rather than zero-masked, so parameter, multiply–accumulate (MAC), memory, and latency reductions are realizable.

Each of the ResNet-50 U-Net, ConvNeXt-V2-Tiny U-Net, and SegFormer-B2 dense models is pruned independently using group magnitude, group Taylor, and GeoBASP importance. The three nominal pruning levels target 25%, 50%, and 70% reductions in dense-model MACs. Using a computation target rather than a fixed channel percentage allows fairer comparison across architectures. Model and threshold selection is confined to the validation partition; the held-out test data are not involved.

3.3.1. Pruning Objective, Units, and Baselines

Let $f(\mathbf{x}; \theta)$ be a fully trained dense segmentation model and let $\mathcal{G} = \{G_1, \dots, G_m\}$ be its candidate structural groups. A group is the set of coupled channels that DepGraph requires to be removed together for the network to remain executable. For a requested

MAC reduction ρ , structured pruning seeks a subset $\mathcal{P} \subset \mathcal{G}$ that retains segmentation performance while satisfying

$$\text{MAC}(f_{\theta \setminus \mathcal{P}}) \leq (1 - \rho) \text{MAC}(f_{\theta}). \quad (1)$$

Because this subset-selection problem is combinatorial, each method assigns an importance score to every valid group and iteratively removes the least important group. Group magnitude uses the mean L2 magnitude of the coupled parameters. Group Taylor accumulates gradients of the mean overall segmentation loss over the calibration patches and scores eligible channels by the sum of absolute weight–gradient products. Contributions from coupled parameterized layers, including eligible LayerNorm weights, are combined by their mean and normalized by the group’s mean score [19].

All gradient-based methods use the same training-only calibration set of 32 complete 512×512 patches from different 20 km blocks. The set balances positive patches by tertiles of total landslide-pixel count, includes difficult negative patches, and covers four geographic strata (northeast, southeast, northwest, and southwest) within the training partition of the Koshi Basin. These strata are defined by the median training-block column and row indices; they are subdivisions of the same study area. Validation and test images are not used to calculate pruning importance.

3.3.2. Proposed GeoBASP Importance

GeoBASP stands for Geography-, Boundary-, and Scar-Aware Structured Pruning. Its central rule is simple: a channel group should be retained when it supports any important part of landslide mapping, even if its average importance is not large. Conversely, a group is suitable for removal when it is consistently unimportant across locations and its removal saves useful computation. The score is constructed in four steps.

Step 1: task-regime sensitivity. The dense weights are kept unchanged, and a temporary unit gate $a_j = 1$ is placed after each eligible output channel j . For calibration block b and task regime r , the first-order sensitivity is

$$T_{jbr} = \left| a_j \frac{\partial \mathcal{L}_{br}}{\partial a_j} \right|. \quad (2)$$

The four regimes are: (i) the balanced focal–Tversky loss on the complete supervised mask; (ii) the same loss within a two-pixel (20 m) morphological boundary band; (iii) the configured segmentation loss on small-scar patches; and (iv) background binary cross-entropy on hard-negative patches. Thus, the criterion measures failure modes that may be diluted in whole-mask loss alone.

Step 2: layer normalization. Raw gradient scales differ between layers. If $\mathcal{C}(j)$ denotes all channels in the layer containing channel j , GeoBASP uses

$$\bar{T}_{jbr} = \frac{T_{jbr}}{|\mathcal{C}(j)|^{-1} \sum_{k \in \mathcal{C}(j)} T_{kbr} + \epsilon}. \quad (3)$$

This makes a score of one approximately equal to the mean sensitivity of its own layer for that block and regime.

Step 3: geographic and regime aggregation. For each regime, normalized scores are aggregated across the available 20 km training blocks $\mathcal{B}_r$ within the Koshi Basin as

$$A_{jr} = \text{median}_{b \in \mathcal{B}_r}(\bar{T}_{jbr}) + \beta \text{MAD}_{b \in \mathcal{B}_r}(\bar{T}_{jbr}), \quad \beta = 1, \quad (4)$$

where MAD is the median absolute deviation. The median represents consistently useful responses; the MAD term protects responses whose importance varies geographically. Equal regime weights are used, and the protected channel score is

$$S_j = \max_{r \in \mathcal{R}} A_{jr}. \quad (5)$$

The maximum is a protection operator: high importance for even one regime is sufficient to retain a channel.

Step 4: dependency-group ranking and MAC utility. For structural group G_i , channel scores are propagated through its dependencies using $S(G_i) = \max_{j \in G_i} S_j$. The analytic MAC saving $\widehat{\Delta M}(G_i)$ sums the contributions of coupled convolutional and linear tensors. In the implementation, savings are normalized over candidate channels within each dependency group to obtain $\widetilde{\Delta M}(G_i)$. The configured sensitivity-to-cost ratio is

$$I_{\text{GeoBASP}}(G_i) = \frac{S(G_i)}{\widetilde{\Delta M}(G_i) + \epsilon}. \quad (6)$$

The ratio is applied channelwise and normalized by its within-group mean before global ranking. A constant within-group cost therefore cancels, leaving the ResNet-50 importance scores and rankings unchanged by MAC utility. The lowest-ranked eligible channels are removed, and achieved whole-model MACs are remeasured after physical removal.

3.3.3. GeoBASP Pruning Procedure

Algorithm 1 summarizes the complete GeoBASP procedure. Importance is recalculated after every structural change because channel dependencies and task-specific sensitivity change as pruning proceeds.

Algorithm 1 GeoBASP structured pruning and recovery.

Require: Converged dense model M ; training-only calibration set $\mathcal{C}$; validation set $\mathcal{V}$; ordered MAC-reduction targets $\mathcal{R}$; recovery length $E = 30$ epochs

Ensure: Recovered pruned model M_ρ for every $\rho \in \mathcal{R}$

```

1: for each  $\rho \in \mathcal{R}$  in ascending order do
2:   repeat
3:     Attach temporary channel gates to eligible layers of  $M$ 
4:     Compute task-specific sensitivities using Equations (2)–(5) on  $\mathcal{C}$ 
5:     Form valid dependency groups and compute  $I_{\text{GeoBASP}}$  using Equation (6) and apply
      within-group mean normalization
6:     Select the lowest-ranked eligible groups using a requested 1% channel-pruning step under
      the common dependency constraints
7:     Physically remove the selected channels while enforcing protected layers and width
      alignment
8:     Remeasure the achieved whole-model MAC reduction  $\widehat{\rho}$ 
9:   until  $\widehat{\rho} \geq \rho$ 
10:   Retain the structural state whose measured reduction is closest to  $\rho$ 
11:   Recover  $M$  for  $E$  epochs and use  $\mathcal{V}$  to retain the checkpoint with the largest validation Dice
12:   Set  $M_\rho \leftarrow M$  and recompute importance on the recovered model
13: end for
14: return  $\{M_\rho : \rho \in \mathcal{R}\}$ 

```

GeoBASP, magnitude, and Taylor use the same DepGraph eligibility rules and global, isomorphic pruning policy. Remaining widths are aligned to multiples of eight; the first input projection and final classifier output are protected. For the two U-Nets, groups couple convolutional channels to dependent normalization and downstream convolution or decoder inputs. In SegFormer-B2, eligible attention-projection, feed-forward, and decoder dimensions are pruned. The attention-head counts remain fixed at 1, 2, 5, and 8 across its

four stages. For grouped attention dimensions, the pruner averages scores across heads and removes the selected within-head positions from every head. Query, key, and value projection widths remain consistent, with dependent output-projection dimensions updated jointly. After each step, the per-head width d_h and attention scale $d_h^{-1/2}$ are refreshed.

Recovery uses AdamW with encoder and decoder learning rates of 10^{-5} and 10^{-4} , respectively. The main method comparison evaluates the validation-selected operating point on held-out test data.

3.3.4. Adaptation of SIRFP

SIRFP uses pairwise spatial redundancy and greedy retained-channel selection on ResNet-50 and ConvNeXt-V2-Tiny U-Nets [23]. The implementation is a post-training adaptation: graph statistics are collected from trained checkpoints rather than during dense training. At each pruning stage, weights are frozen, and unpooled output maps are collected from the same 32 complete 512×512 training calibration patches per seed, with batch size one and no augmentation. ConvNeXt channel-MLP outputs are converted from NHWC to NCHW. Absolute activation magnitudes are normalized spatially into channel probability distributions, using a uniform distribution for zero maps. Redundancy is $\ln 2$ minus Jensen–Shannon divergence, accumulated with a zero-initialized exponential moving average of coefficient 0.99.

Graphs of coupled outputs are averaged within dependency groups and sliced after physical removal. Greedy deletion degrees are recomputed at each micro-step; their negatives rank channels without per-group normalization. SIRFP shares the global, isomorphic pruning policy, protected input/classifier layers, eight-channel alignment and 1% pruning steps. ConvNeXt MLP widths and dependent global-response-normalization parameters are updated jointly. Each architecture starts from the common dense checkpoint and follows progressive 25%, 50% and 70% targets, selecting the nearest measured MAC reduction and recovering for 30 epochs per stage. Recovery uses the shared AdamW and augmentation settings, checkpoint selection by validation Dice and a fixed grid of 21 probability thresholds.

3.3.5. Calibration and Component Analysis

Calibration sensitivity was evaluated on ResNet-50 U-Net using balanced sets of 16, 32 and 64 patches, containing 12/24/48 positives and 4/8/16 negatives. Two further 32-patch conditions resampled hard negatives from a fixed training-only pool or removed positive size/geographic quotas. Three sampling seeds (9026–9028) varied membership within paired families, with nested size sets, a common dense checkpoint and a fixed recovery seed (2026). Each of the 15 runs followed progressive pruning to 25% and 50%, with recalibration between stages and 30 recovery epochs at each stage. Means, sample SD and maxima include all runs and quantify variation due to calibration sampling.

Component analysis used seven sensitivity/aggregation conditions on ResNet-50 at the 50% MAC target. All conditions shared one dense checkpoint, a fixed 32-patch calibration set (24 positive, eight hard-negative) and the same progressive 25%/50% schedule with 30 recovery epochs per stage. To preserve classifier inputs during pruning, the ResNet component analysis protected all 32 final decoder outputs feeding the binary classifier at both pruning transitions; their input channels remained prunable. To examine architecture dependence, ConvNeXt-V2-Tiny U-Net was evaluated using overall, boundary and small-scar sensitivity, with and without protection of its 32 final decoder outputs, under the same calibration and recovery schedule. The main comparison and calibration study used the input/classifier-output protection described above. The calibration and component

analyses selected checkpoints and probability thresholds on validation data; thresholds at the 50% target were applied unchanged to held-out test data.

3.4. Evaluation and Statistical Analysis

Primary accuracy metrics are landslide-class Dice and intersection over union (IoU). Precision, recall, boundary F1, object recall stratified by scar area, and false-positive area per square kilometer on negative patches are secondary metrics. Overall pixel accuracy is not treated as a primary measure because background dominates the images.

Unless otherwise stated, segmentation metrics are reported as mean $\pm$ sample standard deviation across three seeds (2026, 2027 and 2028). The calibration study instead varies sampling with recovery fixed (Section 3.3.5). Pooled architecture summaries and single-checkpoint evaluations are identified separately; latency statistics describe repeated inference sessions.

Efficiency is quantified using parameter count, model size, and MACs. These measures describe storage and arithmetic reduction independently of a specific processor [34]. Batch-one latency is included as a secondary practical measure because MACs and runtime need not change in the same proportion [35]. The selected operating point is the most compressed model whose spatial-block bootstrap 95% confidence interval for $\text{Dice}_{\text{pruned}} - \text{Dice}_{\text{dense}}$ has a lower bound greater than the non-inferiority margin of -0.02 . The margin represents the largest Dice reduction accepted for a useful compact model. The GeoBASP model and matched magnitude, Taylor, and SIRFP models use the selected ConvNeXt seed (2028) and nominal 25% target. Their probability thresholds are refined on validation data in steps of 0.005 and applied unchanged to the held-out test region. Table 2 reports these single checkpoints; Table 3 summarizes the pruning runs. Latency is measured on a laptop with an Intel Core i7-14700HX CPU, NVIDIA GeForce RTX 5060 Laptop GPU (8 GB VRAM), and 32 GB RAM, running Windows 11 and PyTorch 2.13 with CUDA 13.0. All models use CUDA Graph replay in evaluation mode with gradients disabled, strict FP32, TF32 disabled, and cuDNN autotuning enabled. Batch-one $1 \times 4 \times 512 \times 512$ inputs are GPU-resident. Three independent measurement sessions, each in a separate process, use 50 warm-up and 200 timed inference iterations. Reported latency is the mean and sample SD of the three session medians.

Table 2. Held-out ConvNeXt-V2-Tiny performance near 25% MAC reduction. Segmentation metrics report single checkpoints, with validation thresholds refined in steps of 0.005; Table 3 reports aggregate results. Intervals compare pruned and dense Dice over paired 20 km blocks. Latency is CUDA Graph FP32 mean $\pm$ SD (ms) across three sessions on the RTX 5060 Laptop GPU. Bold identifies the best reported pruned value; achieved MAC reduction is not ranked.

Method	ΔMAC (%)	Params (M)	MACs (G)	Val. Dice	Test Dice	Test IoU	ΔDice 95% CI	NI	Latency (ms)
Dense	0	34.957	53.009	0.4873	0.5159	0.3477	–	–	31.31 $\pm$ 0.13
Magnitude [16]	24.09	29.390	40.238	0.3058	0.3479	0.2106	[−0.2060, −0.1131]	No	24.58 $\pm$ 0.04
Taylor [19]	25.80	28.409	39.330	0.4620	0.4740	0.3106	[−0.0666, −0.0290]	No	24.85 $\pm$ 0.02
GeoBASP (Ours)	24.22	30.648	40.171	0.4848	0.5054	0.3382	[− 0.0191 , − 0.0033]	Yes	25.58 $\pm$ 0.04
SIRFP [23]	24.31	25.499	40.124	0.4580	0.5087	0.3412	[−0.0246, 0.0073]	No	26.43 $\pm$ 0.18

Table 3. Validation Dice after structured pruning. Overall pools nine architecture–run values for magnitude, Taylor, and GeoBASP; † marks SIRFP’s six values from the two U-Nets, with SegFormer unevaluated (–). Achieved MAC reduction uses the same coverage. NI counts intervals above the –0.02 margin, out of nine or six values. Bold identifies the highest Dice per architecture and target; overall values are not ranked because coverage differs.

Method	Target Δ MAC	Achieved Δ MAC	ResNet-50 Dice	ConvNeXt-V2 Dice	SegFormer-B2 Dice	Overall Dice	NI Passes
Dense	0%	0%	0.4878	0.4873	0.4413	0.4721	–
Magnitude [16]	25%	24.79%	0.3399 ± 0.0050	0.3031 ± 0.0096	0.3928 ± 0.0030	0.3452 ± 0.0394	0/9
	50%	49.70%	0.2087 ± 0.0214	0.2165 ± 0.0097	0.3472 ± 0.0020	0.2575 ± 0.0684	0/9
	70%	70.10%	0.1776 ± 0.0102	0.2132 ± 0.0050	0.3543 ± 0.0014	0.2484 ± 0.0811	0/9
Taylor [19]	25%	25.54%	0.4637 ± 0.0029	0.4632 ± 0.0071	0.4003 ± 0.0020	0.4424 ± 0.0318	0/9
	50%	49.81%	0.4583 ± 0.0103	0.3344 ± 0.1507	0.3577 ± 0.0041	0.3835 ± 0.0947	0/9
	70%	70.07%	0.4274 ± 0.0162	0.2997 ± 0.2084	0.3513 ± 0.0047	0.3595 ± 0.1184	0/9
GeoBASP (Ours)	25%	24.36%	0.4736 ± 0.0054	0.4777 ± 0.0054	0.3940 ± 0.0054	0.4484 ± 0.0411	2/9
	50%	50.04%	0.4635 ± 0.0112	0.4612 ± 0.0076	0.3520 ± 0.0040	0.4256 ± 0.0556	0/9
	70%	70.01%	0.4154 ± 0.0023	0.4192 ± 0.0039	0.3507 ± 0.0021	0.3951 ± 0.0334	0/9
SIRFP [23]	25%	24.88%	0.4580 ± 0.0024	0.4561 ± 0.0033	–	0.4570 ± 0.0028 †	0/6
	50%	49.71%	0.4231 ± 0.0044	0.4313 ± 0.0098	–	0.4272 ± 0.0081 †	0/6
	70%	70.04%	0.3876 ± 0.0013	0.4315 ± 0.0041	–	0.4095 ± 0.0242 †	0/6

4. Results

The results progress from dense segmentation performance to compact-model analysis. First, the input and training ablation is followed by a comparison of the three dense architectures, establishing the reference models for pruning. The pruning comparison then evaluates GeoBASP, magnitude, Taylor and SIRFP at three MAC-reduction targets. Finally, component ablations, qualitative maps, scar-size analysis, and sensitivity tests are used to examine why the proposed method succeeds or fails under different conditions. Validation results guide model and threshold selection, while the held-out test partition measures within-basin spatial generalization.

4.1. Dense Landslide Segmentation

Before comparing architectures and pruning methods, the dense input and training choices were examined through a sequential ablation using ConvNeXt-V2-Tiny U-Net. The first configuration used the three spectral bands B11/B08/B04 with empirical sampling and BCE–Dice loss. The second retained the same spectral input but introduced foreground-aware sampling and focal–Tversky loss. The proposed configuration then added terrain slope as a fourth channel while keeping the imbalance-aware training strategy fixed. This design separates the influence of imbalance handling from the contribution of terrain information. Table 4 summarizes the validation and geographically separated test results.

Table 4. Effect of spectral, imbalance-aware, and terrain-aware configurations on ConvNeXt-V2-Tiny U-Net. Probability thresholds were determined from validation data.

Configuration	Channels	Training Strategy	Val. Dice	Test Dice	Test IoU
Standard spectral	B11/B08/B04	Empirical sampling; BCE–Dice	0.4369 ± 0.0033	0.5010 ± 0.0086	0.3343 ± 0.0076
Imbalance-aware spectral	B11/B08/B04	Foreground-aware; focal–Tversky	0.4390 ± 0.0022	0.4710 ± 0.0067	0.3081 ± 0.0057
Terrain-aware (proposed)	B11/B08/B04/slope	Foreground-aware; focal–Tversky	0.4855 ± 0.0018	0.5202 ± 0.0037	0.3515 ± 0.0034

Note: Bold values indicate the highest result in each metric column.

The comparison shows that terrain information, rather than imbalance handling alone, produced the main improvement. Compared with the standard spectral model, imbalance-

aware sampling and loss changed mean validation Dice only from 0.4369 to 0.4390 and reduced mean test Dice from 0.5010 to 0.4710. It also increased false positives on negative test patches from 1.42×10^{-5} to 2.90×10^{-4} of supervised background pixels. Adding slope to this otherwise unchanged configuration raised mean validation Dice to 0.4855 and mean test Dice to 0.5202. The matched terrain-minus-imbalance test improvement was 0.0492 ± 0.0068 Dice, and the terrain-aware model exceeded the standard spectral model in every run. Thus, targeted sampling and asymmetric loss helped expose rare foreground during optimization, but the added physical terrain cue was required to improve held-out performance within the basin.

For the matched architecture comparison, all three networks used a single common run and the same terrain-aware configuration. ConvNeXt-V2-Tiny U-Net achieved the highest performance on the geographically separated test region, with Dice 0.5159 and IoU 0.3477, compared with 0.4869 and 0.3218 for ResNet-50 U-Net and 0.4651 and 0.3030 for SegFormer-B2. The corresponding precision–recall pairs were 0.5609–0.4776, 0.5679–0.4261, and 0.5125–0.4258. The 20 km-block bootstrap Dice intervals were 0.4373–0.5646, 0.4024–0.5401, and 0.3745–0.5252, respectively. Their overlap shows that the architecture ranking is less certain between geographic blocks than the global values alone suggest.

Performance varied with scar size. ConvNeXt-V2-Tiny recovered the largest pixel fraction for scars up to 100 pixels, whereas SegFormer-B2 was strongest above 100 pixels and gave the smallest false-positive fraction on negative patches. ConvNeXt-V2-Tiny was therefore used for the detailed dense-model and compact-model analyses. Figure 3 provides representative held-out predictions from this dense reference model, including successful, partial, missed-scar, and difficult-negative cases.

4.2. Effect of Structured Pruning

Table 3 compares GeoBASP, magnitude, Taylor, and SIRFP on validation data. The first three methods cover all three architectures; SIRFP covers the two U-Nets. Among methods evaluated on all three architectures, GeoBASP has the highest overall mean Dice at 25%, 50%, and 70% MAC reduction: 0.4484, 0.4256, and 0.3951. SIRFP's overall values average its two evaluated architectures, so comparisons with GeoBASP use the corresponding architecture columns or the paired differences in Figure 4.

GeoBASP leads ConvNeXt-V2-Tiny at 25% and 50%, with Dice of 0.4777 ± 0.0054 and 0.4612 ± 0.0076 ; SIRFP leads at 70% with 0.4315 ± 0.0041 vs. 0.4192 ± 0.0039 . GeoBASP also leads ResNet-50 at 25% and 50%, whereas Taylor leads there at 70% and for SegFormer-B2 at 25% and 50%. Magnitude pruning causes substantial accuracy losses for both U-Nets.

At the 70% target, SIRFP retains 5.88 million parameters on ResNet-50 and 7.48 million on ConvNeXt-V2-Tiny, illustrating the storage–accuracy trade-off at similar arithmetic budgets.

Two GeoBASP results at the 25% target satisfied the non-inferiority criterion; none of the magnitude, Taylor, or SIRFP results did. No method met this criterion at the two higher reduction levels. Figure 4 compares all four methods over the two U-Nets, shows ConvNeXt-V2-Tiny behaviour, and presents GeoBASP's paired differences over the shared architectures for each comparison.

4.3. Accuracy–Efficiency Trade-Off

At the selected 25% MAC-reduction operating point, GeoBASP reaches test Dice 0.5054 vs. dense Dice 0.5159 (Table 2). Its paired interval $[-0.0191, -0.0033]$ passes the -0.02 margin. SIRFP has a higher global test Dice (0.5087), but its interval $[-0.0246, 0.0073]$ fails non-inferiority, as do magnitude and Taylor. GeoBASP and SIRFP have similar boundary F1, 0.6061 and 0.6054, respectively.

GeoBASP removed 24.2% of MACs and 12.3% of parameters at this operating point. At the 50% and 70% targets, representative ConvNeXt-V2-Tiny GeoBASP models removed 33.3% and 53.7% of parameters, respectively. These reductions directly lower arithmetic work and model storage. The mean within-session latency reduction was 18.3% with CUDA Graph FP32 (31.31 to 25.58 ms), below the 24.2% MAC reduction. Practical runtime therefore depends on the execution library and hardware as well as on MAC count.

4.4. Understanding the GeoBASP Components

Final-feature protection provides a structural safeguard against accuracy degradation during pruning. Without this safeguard, low precision in ResNet-50 was associated with removal of every positively weighted classifier input. Since final ReLU features are nonnegative, an all-negative classifier bounds foreground probability by the sigmoid of its bias. The safeguard preserves these input channels.

With final-feature protection, ResNet-50 validation Dice at 50% was 0.4390–0.4680, with mean precision 0.3812–0.4138 and recall 0.4966–0.5571. Adding small-scar sensitivity after boundary sensitivity improved test Dice in every paired comparison, by a mean of 0.0146. Full GeoBASP exceeded the mean test Dice of overall sensitivity by 0.0217 and attained the highest individual test Dice (0.4881); the small-scar variant had the highest mean Dice (Table 5). Geographic aggregation did not consistently improve mean Dice, although four-regime mean worst-positive-block validation Dice increased from 0.1259 to 0.1641. At all 42 calibration points, importance scores and rankings were identical with and without MAC utility. Component rankings are specific to ResNet-50 at the 50% target. Since final-feature protection also changes channel allocation, these comparisons do not isolate its effect.

At the 50% target, ConvNeXt-V2-Tiny’s overall + boundary + small-scar variant achieved a validation Dice of 0.4404 ± 0.0092 and a test Dice of 0.4875 ± 0.0025 without final-feature protection; at 48.83%, it achieved MAC reduction. With protection, the corresponding Dice scores were 0.4479 ± 0.0076 and 0.4800 ± 0.0055 , at 49.80% reduction. Both conditions retained positively weighted classifier inputs and produced consistent Dice; protection did not uniformly improve accuracy.

Table 5. GeoBASP component analysis on ResNet-50 at the 50% MAC target. All seven variants share the fixed 32-patch calibration set, progressive schedule and final-feature safeguard. Δ MAC is the mean achieved reduction; maxima are descriptive and may come from different runs on each split. Bold identifies the highest mean or maximum in each Dice column.

Variant	Δ MAC (%)	Val Dice	Val Max	Test Dice	Test Max
Overall sensitivity	49.73	0.4568 ± 0.0032	0.4598	0.4383 ± 0.0215	0.4594
+ boundary	49.82	0.4538 ± 0.0034	0.4559	0.4507 ± 0.0099	0.4598
+ small scar	49.77	0.4575 ± 0.0109	0.4680	0.4653 ± 0.0042	0.4699
All regimes, no geography	50.25	0.4530 ± 0.0132	0.4652	0.4613 ± 0.0075	0.4686
+ geographic median-MAD	49.92	0.4526 ± 0.0083	0.4592	0.4420 ± 0.0295	0.4678
Full GeoBASP (+ MAC utility)	50.20	0.4536 ± 0.0085	0.4629	0.4599 ± 0.0263	0.4881
Overall sensitivity + geographic median-MAD	50.07	0.4483 ± 0.0032	0.4513	0.4367 ± 0.0437	0.4646

4.5. Mapping Behaviour and Sensitivity

Figure 3 complements the aggregate metrics with representative held-out predictions. In the first example, the dense model followed most of the reference scar and achieved a patch Dice of 0.769. The partially detected example reached 0.492, whereas a small 116-pixel scar was missed. The difficult negative patch produced 52 false-positive pixels. These examples illustrate characteristic successes and errors; population-level performance is given by the quantitative results.

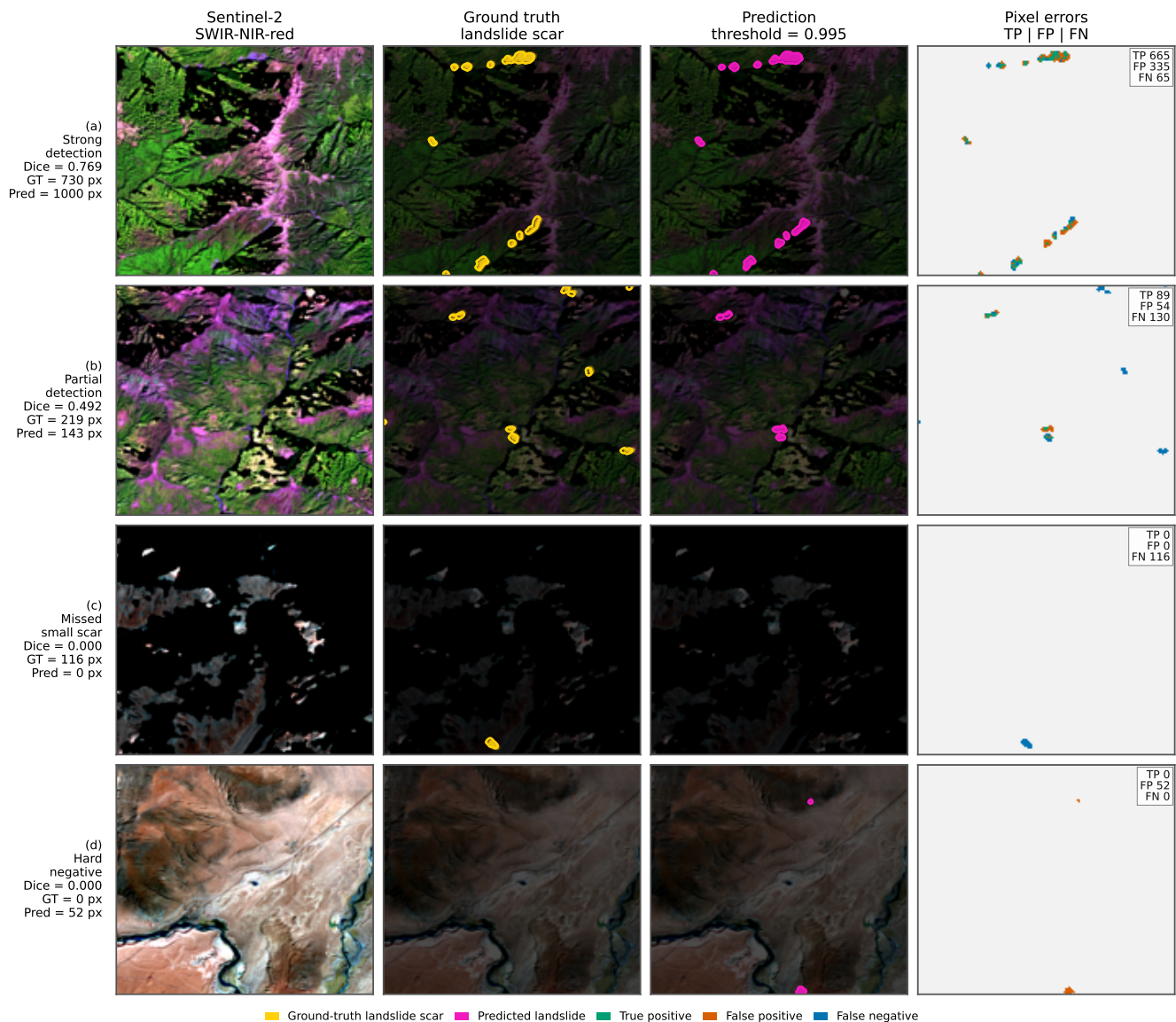


Figure 3. Representative held-out predictions from the highest-performing dense model, ConvNeXt-V2-Tiny U-Net. Columns show the Sentinel-2 composite, inventory-derived landslide scars in yellow, predictions in magenta, and pixel errors (true positive, green; false positive, orange; false negative, blue) on a light categorical background. Because the masks are sparse, the ground-truth, prediction, and error overlays are expanded outward by four raster pixels for display only; the printed counts and all metrics use the original, unexpanded masks. (a) Well-delineated scar; (b) partial detection; (c) missed small scar; and (d) difficult negative patch.

Figure 5 shows how increasing compression affects the same held-out patch presented in the first row of Figure 3. The dense, 25%, 50%, and 70% models obtained patch Dice of 0.769, 0.767, 0.767, and 0.765, respectively, and all retained the main scar structure. This example shows that useful features can remain after strong pruning in an individual image. However, the aggregate results in Table 3 show a clearer accuracy loss and greater uncertainty at high compression.

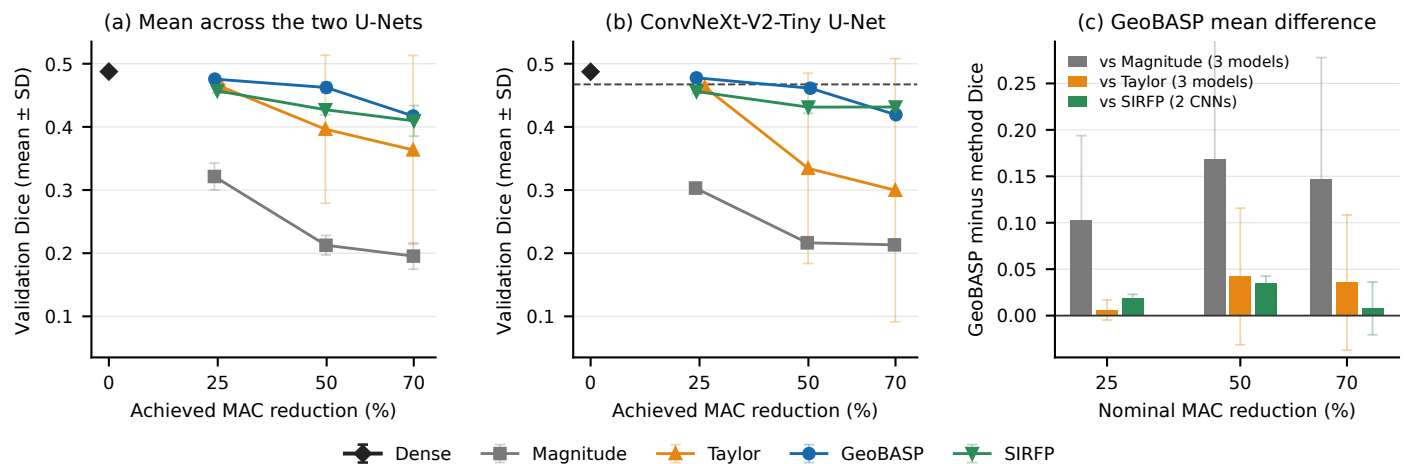


Figure 4. Validation accuracy across MAC-reduction targets. (a) Mean Dice over ResNet-50 and ConvNeXt-V2-Tiny U-Nets, pooling their six runs for every method. (b) ConvNeXt-V2-Tiny mean Dice, with a dashed line 0.02 below dense Dice. (c) GeoBASP-minus-method Dice, first averaged within each architecture and then over shared architectures: three for magnitude/Taylor and two for SIRFP. Positive values favour GeoBASP. Error bars show sample SD across Dice values (a,b), dense models (a), and architecture differences (c).

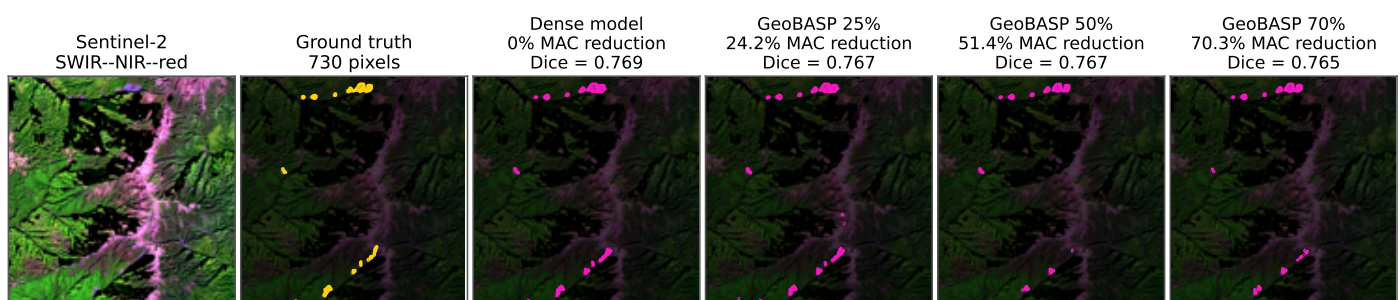


Figure 5. A representative held-out example across GeoBASP compression levels. Yellow marks the inventory-derived landslide scar and magenta marks model predictions. The same patch is shown in the first row of Figure 3. It illustrates the preservation of the main scar structure at 25%, 50%, and 70% nominal MAC reduction; aggregate performance is reported in Table 3.

Using the stricter condition that at least half of each reference component must be recovered, the dense ConvNeXt-V2-Tiny detected 11.8%, 35.7%, 52.2%, and 66.7% of test scars in the 1–9, 10–25, 26–100, and above-100-pixel groups. The corresponding GeoBASP model obtained 12.3%, 34.6%, 51.8%, and 63.0%. Moderate pruning therefore preserved the response across scar sizes. Both models remained weak for the smallest scars, where a few pixels determine the object result.

Removing the reconstructed 4500 m supervision limit added 169,881 validation pixels. Dense Dice changed from 0.4876 to 0.4849 and GeoBASP Dice from 0.4849 to 0.4830. The paired GeoBASP-minus-dense interval remained above the non-inferiority margin. The pruning conclusion was therefore robust to the elevation rule used to reconstruct the supervision domain.

Calibration-Set Sensitivity

Table 6 shows similar validation Dice across the five conditions at 25% pruning, but substantial calibration dependence at 50%. One balanced-16 selection yielded validation Dice 0.1944 and test Dice 0.0629; one balanced-32 selection yielded 0.2694 and 0.0596, respectively, despite much higher maxima in both conditions. Every balanced-64 selection avoided these failures. Resampling hard negatives also gave consistent test Dice,

whereas removing positive quotas increased test variation despite stable validation results. Thus, neither a 32-patch calibration set nor balancing alone guarantees robustness at the higher target.

Table 6. ResNet-50 calibration sensitivity under progressive 25%/50% MAC pruning. Variation reflects calibration sampling with recovery fixed; maxima are descriptive and may come from different runs on each split. Balanced 16/32/64 use 12/24/48 positive and 4/8/16 negative patches. Both composition variants use 24 positives and eight negatives; the unstratified-positive variant shares the resampled negatives. Test evaluation uses the validation-selected thresholds unchanged.

Calibration	25% Val. Dice	50% Val. Dice	Val. Max	50% Test Dice	Test Max
Balanced 16	0.4749 ± 0.0059	0.3717 ± 0.1536	0.4621	0.3352 ± 0.2359	0.4785
Balanced 32	0.4737 ± 0.0061	0.3933 ± 0.1074	0.4604	0.3399 ± 0.2431	0.4929
Balanced 64	0.4781 ± 0.0070	0.4696 ± 0.0061	0.4767	0.4709 ± 0.0092	0.4784
32, resampled negatives	0.4690 ± 0.0048	0.4573 ± 0.0120	0.4709	0.4653 ± 0.0019	0.4665
32, unstratified positives	0.4757 ± 0.0049	0.4561 ± 0.0028	0.4587	0.4310 ± 0.0354	0.4683

5. Discussion

The evidence supports GeoBASP most clearly at moderate compression: only GeoBASP met non-inferiority at the 25% MAC target, including the selected ConvNeXt-V2-Tiny operating point on held-out test data (Tables 2 and 3). At 50% and 70%, favourable relative rankings still involved greater accuracy loss. A feature that is infrequent across the whole dataset may still be essential for a small scar, an uncertain boundary, or a particular geographic subregion. Task-and-block aggregation is designed to retain such features, but its advantage depends on architecture, recovery and pruning budget [20,21].

SegFormer’s pruning dynamics help explain the different rankings of Taylor and GeoBASP. Convolutional removal changes feature maps and dependent normalization and inputs; query/key removal changes attention dot products, while value removal changes the representation passed to the output projection [15]. The head count remains fixed, so pruning removes corresponding coordinates from every head. In the first block of the second stage at 50% MAC reduction (seed 2026), both methods retained 16 query/key/value dimensions (eight per head), down from 128, but Taylor retained 416 feed-forward dimensions and GeoBASP 376, down from 512. Their loss-based rankings can therefore produce equal attention widths with different feed-forward structures. Both are first-order criteria: the observed Dice gap also reflects structural coupling, achieved budgets and recovery; this structural comparison does not isolate the cause of the Dice difference. A controlled attention/feed-forward ablation is needed for causal attribution.

MAC and parameter reductions provide hardware-independent measures of arithmetic demand and model storage. The smaller latency gain (Table 2) is consistent with execution costs that remain after pruning. Feature-map access, normalization, interpolation and skip concatenation may be limited by GPU memory bandwidth [34,36]. CUDA Graph replay does not eliminate feature-map traffic or per-forward synchronization. Changed widths can alter kernel selection, tiling and parallelism; eight-channel alignment does not guarantee equal utilization [35,37]. These are plausible explanations, since bandwidth, occupancy and kernel timings were not profiled. The measurements establish a device- and backend-specific gain; identifying the dominant costs requires profiling.

The dense ablation also provides information about landslide mapping in Nepal. Foreground-aware sampling and asymmetric loss alone did not improve the geographically separated test result and increased false alarms on negative patches. In contrast, adding slope to the same training configuration improved test Dice consistently across runs. Slope supplies a direct physical cue while the spectral bands distinguish vegetation, moisture, and exposed soil.

Pixel Dice measures agreement with inventory masks and does not by itself establish event-level completeness or operational reliability. The omissions and false alarms in Section 4.5 remain consequential. The outputs could support candidate-scar screening with expert verification against independent imagery or field evidence, but suitability for standalone warnings or complete disaster-response maps is unestablished. Low small-scar recall means unflagged areas cannot be assumed landslide-free. Pruning non-inferiority limits loss relative to the dense model without validating an operational workflow, which requires evaluation of new-event detection, omissions, false alarms, imagery availability and time to a verified map.

Spatial resolution further limits small-scar detection. Sentinel-2 B04/B08 are natively 10 m and B11 20 m [3], while slope comes from a 30 m DEM; resampling to 10 m cannot recover finer detail. Mixed pixels reduce scar-background contrast, and rasterization and boundary uncertainty disproportionately affect small objects. Poor 1–9-pixel recall (Section 4.5) is consistent with these constraints, although the contribution of resolution was not isolated experimentally. Higher-resolution imagery and labels warrant evaluation.

The comparisons concern the evaluated implementations. SIRFP covers the two U-Nets; its published pipeline, CAP and SIRFP on SegFormer-B2 remain unevaluated. Software and runtime environments were not fully controlled across comparisons, and the rankings do not establish statistical superiority. Calibration membership can cause severe failures at 50% pruning (Table 6); the sensitivity study fixes the dense checkpoint and recovery seed, leaving broader architecture and recovery-seed robustness untested. Component rankings remain specific to the protected ResNet-50 setting.

Geographic conclusions concern within-basin spatial generalization. GeoBASP's scores use training blocks from one transboundary Koshi Basin inventory; the 20 km guard columns control local leakage without testing transfer to an independent region. The inventory was prepared using model-assisted mapping followed by manual editing and is therefore not equivalent to an independent set of field observations. The exact systematic mapping footprint was not included in the dataset and was reconstructed from the available information, although the elevation sensitivity test did not change the conclusion. Evaluation on geographically distinct datasets is needed to assess transfer under different terrain, sensor and annotation conditions.

6. Conclusions

GeoBASP was developed to prune fully trained landslide-segmentation networks while retaining channels needed for scars, boundaries, difficult backgrounds, and particular geographic areas. The terrain-aware dense model achieved a mean test Dice of 0.5202 ± 0.0037 on the spatially separated test region. A matched comparison showed that adding slope improved this result beyond imbalance-aware spectral training alone.

Among methods evaluated across all three architectures, GeoBASP achieved the highest overall mean validation Dice at 25%, 50% and 70% MAC reduction. No method met non-inferiority at 50% or 70%. For ConvNeXt-V2-Tiny U-Net, the selected model reduced MACs by 24.2% and parameters by 12.3%, retained a test Dice of 0.5054 vs. 0.5159 for the dense model, and met the predefined non-inferiority criterion. These findings support moderate compression within the evaluated basin; stronger pruning incurs greater accuracy loss.

GeoBASP's advantage depends on architecture and budget. It exceeds SIRFP's mean validation Dice at all ResNet-50 targets and at 25% and 50% for ConvNeXt-V2-Tiny, whereas SIRFP leads ConvNeXt at 70%. Taylor also outperforms GeoBASP in some settings. SIRFP's higher global test Dice at the selected operating point does not meet spatial-block non-inferiority. Because reduced MACs did not translate proportionally into shorter runtime,

deployment should also be assessed on the intended hardware. Cross-region transfer and broader calibration and component robustness require further evaluation.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The ICIMOD Landslide Inventory of Koshi Basin 2024 is available at <https://doi.org/10.26066/rds.1973522> under CC BY 4.0. Sentinel-2 Level-2A imagery and the Copernicus DEM GLO-30 are publicly available through the Copernicus Data Space Ecosystem. Processing code, spatial split definitions, and trained models are publicly available on GitHub (v1.0.0) at <https://github.com/ghimiredhikura/GeoBASP/releases/tag/v1.0.0> (accessed on 15 October 2026). Derived results can be regenerated with the released code and data.

Acknowledgments: During preparation of this study, the author used OpenAI ChatGPT (GPT-5 and 6 series; accessed June–October 2026) for code, data, and language assistance. The author reviewed and edited all outputs and takes full responsibility for the content of this publication.

Conflicts of Interest: The author declares no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

AOI	Area of interest
CNN	Convolutional neural network
DEM	Digital elevation model
GT	Ground truth
IoU	Intersection over union
MAC	Multiply–accumulate operation
MGRS	Military Grid Reference System
NIR	Near infrared
SCL	Scene Classification Layer
SWIR	Short-wave infrared
UTM	Universal Transverse Mercator

References

1. Guzzetti, F.; Mondini, A.C.; Cardinali, M.; Fiorucci, F.; Santangelo, M.; Chang, K.T. Landslide Inventory Maps: New Tools for an Old Problem. *Earth-Sci. Rev.* **2012**, *112*, 42–66. <https://doi.org/10.1016/j.earscirev.2012.02.001>.
2. Williams, J.G.; Rosser, N.J.; Kinsey, M.E.; Benjamin, J.; Oven, K.J.; Densmore, A.L.; Milledge, D.G.; Robinson, T.R.; Jordan, C.A.; Dijkstra, T.A. Satellite-Based Emergency Mapping Using Optical Imagery: Experience and Reflections from the 2015 Nepal Earthquakes. *Nat. Hazards Earth Syst. Sci.* **2018**, *18*, 185–205. <https://doi.org/10.5194/nhess-18-185-2018>.
3. Drusch, M.; Del Bello, U.; Carlier, S.; Colin, O.; Fernandez, V.; Gascon, F.; Hoersch, B.; Isola, C.; Laberinti, P.; Martimort, P.; et al. Sentinel-2: ESA’s Optical High-Resolution Mission for GMES Operational Services. *Remote Sens. Environ.* **2012**, *120*, 25–36. <https://doi.org/10.1016/j.rse.2011.11.026>.
4. Stumpf, A.; Kerle, N. Object-Oriented Mapping of Landslides Using Random Forests. *Remote Sens. Environ.* **2011**, *115*, 2564–2577. <https://doi.org/10.1016/j.rse.2011.05.013>.
5. Prakash, N.; Manconi, A.; Loew, S. Mapping Landslides on EO Data: Performance of Deep Learning Models vs. Traditional Machine Learning Models. *Remote Sens.* **2020**, *12*, 346. <https://doi.org/10.3390/rs12030346>.
6. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, Munich, Germany, 5–9 October 2015; Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F., Eds.; Lecture Notes in Computer Science; Volume 9351, pp. 234–241. https://doi.org/10.1007/978-3-319-24574-4_28.
7. Chen, L.C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder–Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In Proceedings of the Computer Vision—ECCV 2018, Munich, Germany, 8–14 September 2018; Ferrari,

- V., Hebert, M., Sminchisescu, C., Weiss, Y., Eds.; Lecture Notes in Computer Science; Volume 11211, pp. 833–851. https://doi.org/10.1007/978-3-030-01234-2_49.
8. Yu, B.; Chen, F.; Xu, C. Landslide Detection Based on Contour-Based Deep Learning Framework in Case of National Scale of Nepal in 2015. *Comput. Geosci.* **2020**, *135*, 104388. <https://doi.org/10.1016/j.cageo.2019.104388>.
9. Bragagnolo, L.; Rezende, L.R.; da Silva, R.V.; Grzybowski, J.M.V. Convolutional Neural Networks Applied to Semantic Segmentation of Landslide Scars. *CATENA* **2021**, *201*, 105189. <https://doi.org/10.1016/j.catena.2021.105189>.
10. Shahabi, H.; Rahimzad, M.; Tavakkoli Piralilou, S.; Ghorbanzadeh, O.; Homayouni, S.; Blaschke, T.; Lim, S.; Ghamisi, P. Unsupervised Deep Learning for Landslide Detection from Multispectral Sentinel-2 Imagery. *Remote Sens.* **2021**, *13*, 4698. <https://doi.org/10.3390/rs13224698>.
11. Ghorbanzadeh, O.; Meena, S.R.; Shahabi Sorman Abadi, H.; Tavakkoli Piralilou, S.; Lv, Z.; Blaschke, T. Landslide Mapping Using Two Main Deep-Learning Convolution Neural Network Streams Combined by the Dempster–Shafer Model. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **2021**, *14*, 452–463. <https://doi.org/10.1109/JSTARS.2020.3043836>.
12. Ghorbanzadeh, O.; Xu, Y.; Ghamisi, P.; Kopp, M.; Kreil, D.P. Landslide4Sense: Reference Benchmark Data and Deep Learning Models for Landslide Detection. *IEEE Trans. Geosci. Remote Sens.* **2022**, *60*, 5633017. <https://doi.org/10.1109/TGRS.2022.3215209>.
13. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. <https://doi.org/10.1109/CVPR.2016.90>.
14. Woo, S.; Debnath, S.; Hu, R.; Chen, X.; Liu, Z.; Kweon, I.S.; Xie, S. ConvNeXt V2: Co-Designing and Scaling ConvNets with Masked Autoencoders. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 16133–16142. <https://doi.org/10.1109/CVPR52729.2023.01548>.
15. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In Proceedings of the Advances in Neural Information Processing Systems, Virtual Conference, 6–14 December 2021; Volume 34, pp. 12077–12090.
16. Li, H.; Kadav, A.; Durdanovic, I.; Samet, H.; Graf, H.P. Pruning Filters for Efficient ConvNets. In Proceedings of the International Conference on Learning Representations, Toulon, France, 24–26 April 2017.
17. Liu, Z.; Li, J.; Shen, Z.; Huang, G.; Yan, S.; Zhang, C. Learning Efficient Convolutional Networks through Network Slimming. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2736–2744. <https://doi.org/10.1109/ICCV.2017.298>.
18. He, Y.; Zhang, X.; Sun, J. Channel Pruning for Accelerating Very Deep Neural Networks. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 1389–1397. <https://doi.org/10.1109/ICCV.2017.155>.
19. Molchanov, P.; Mallya, A.; Tyree, S.; Frosio, I.; Kautz, J. Importance Estimation for Neural Network Pruning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 11264–11272. <https://doi.org/10.1109/CVPR.2019.01152>.
20. Blalock, D.; Gonzalez Ortiz, J.J.; Frankle, J.; Gutttag, J. What Is the State of Neural Network Pruning? In Proceedings of the 3rd Conference on Machine Learning and Systems (MLSys 2020), Austin, TX, USA, 2–4 March 2020; Volume 2, pp. 129–146.
21. Hoefler, T.; Alistarh, D.; Ben-Nun, T.; Dryden, N.; Peste, A. Sparsity in Deep Learning: Pruning and Growth for Efficient Inference and Training in Neural Networks. *J. Mach. Learn. Res.* **2021**, *22*, 1–124.
22. He, W.; Wu, M.; Liang, M.; Lam, S.K. CAP: Context-Aware Pruning for Semantic Segmentation. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 3–8 January 2021; pp. 960–969. <https://doi.org/10.1109/WACV48630.2021.00100>.
23. Wu, D.; Guo, Z.; Yu, L.; Sang, N.; Gao, C. Structural Pruning via Spatial-Aware Information Redundancy for Semantic Segmentation. *Proc. AAAI Conf. Artif. Intell.* **2025**, *39*, 8368–8376. <https://doi.org/10.1609/aaai.v39i8.32903>.
24. Ghimire, D.; Kim, S.H. Magnitude and Similarity Based Variable Rate Filter Pruning for Efficient Convolution Neural Networks. *Appl. Sci.* **2023**, *13*, 316. <https://doi.org/10.3390/app13010316>.
25. Ghimire, D.; Lee, K.; Kim, S.H. Loss-Aware Automatic Selection of Structured Pruning Criteria for Deep Neural Network Acceleration. *Image Vis. Comput.* **2023**, *136*, 104745. <https://doi.org/10.1016/j.imavis.2023.104745>.
26. Ghimire, D.; Kil, D.; Jeong, S.; Park, J.; Kim, S.H. One-Cycle Structured Pruning via Stability-Driven Subnetwork Search. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Tucson, AZ, USA, 6–10 March 2026; pp. 5467–5476. <https://doi.org/10.1109/WACV61042.2026.00530>.
27. Fang, G.; Ma, X.; Song, M.; Mi, M.B.; Wang, X. DepGraph: Towards Any Structural Pruning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 17–24 June 2023; pp. 16091–16101. <https://doi.org/10.1109/CVPR52729.2023.01544>.
28. ICIMOD. Landslide Inventory of Koshi Basin 2024, 2026. Dataset. Available online: <https://rds.icimod.org/metadata/af73da0a-885b-459d-95ba-2ea0662a7e7c> (accessed on 15 October 2026). <https://doi.org/10.26066/rds.1973522>.

29. Roberts, D.R.; Bahn, V.; Ciuti, S.; Boyce, M.S.; Elith, J.; Guillera-Arroita, G.; Hauenstein, S.; Lahoz-Monfort, J.J.; Schröder, B.; Thuiller, W.; et al. Cross-Validation Strategies for Data with Temporal, Spatial, Hierarchical, or Phylogenetic Structure. *Ecography* **2017**, *40*, 913–929. <https://doi.org/10.1111/ecog.02881>.
30. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal Loss for Dense Object Detection. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2980–2988. <https://doi.org/10.1109/ICCV.2017.324>.
31. Salehi, S.S.M.; Erdogmus, D.; Gholipour, A. Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks. In Proceedings of the Machine Learning in Medical Imaging, Quebec City, QC, Canada, 10 September 2017; Wang, Q., Shi, Y., Suk, H.I., Suzuki, K., Eds.; Lecture Notes in Computer Science; Volume 10541, pp. 379–387. https://doi.org/10.1007/978-3-319-67389-9_44.
32. Kervadec, H.; Bouchtiba, J.; Desrosiers, C.; Granger, E.; Dolz, J.; Ben Ayed, I. Boundary Loss for Highly Unbalanced Segmentation. In Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning, London, UK, 8–10 July, 2019; Proceedings of Machine Learning Research; Volume 102, pp. 285–296.
33. Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Proceedings of the Advances in Neural Information Processing Systems, Vancouver, BC, Canada, 8–14 December 2019; Volume 32, pp. 8024–8035.
34. Sze, V.; Chen, Y.H.; Yang, T.J.; Emer, J.S. Efficient Processing of Deep Neural Networks: A Tutorial and Survey. *Proc. IEEE* **2017**, *105*, 2295–2329. <https://doi.org/10.1109/JPROC.2017.2761740>.
35. Dehghani, M.; Arnab, A.; Beyer, L.; Vaswani, A.; Tay, Y. The Efficiency Misnomer. In Proceedings of the International Conference on Learning Representations, Virtual Conference, 25–29 April 2022.
36. NVIDIA. GPU Performance Background User’s Guide. Available online: <https://docs.nvidia.com/deeplearning/performance/dl-performance-gpu-background/index.html> (accessed on 7 October 2026).
37. NVIDIA. Convolutional Layers User’s Guide. Available online: <https://docs.nvidia.com/deeplearning/performance/dl-performance-convolutional/index.html> (accessed on 7 October 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.