

UC Berkeley

UC Berkeley Previously Published Works

Title

Multiclass Cloud Detection for Very High-Resolution Satellite Imagery Using U-Net: A Multi-Sensor Validation Study

Permalink

<https://escholarship.org/uc/item/2w12z7b5>

Journal

IEEE Access, 14

ISSN

2169-3536

Authors

Sudiana, Dodi

Candra, Danang Surya

Rizkinia, Mia

et al.

Publication Date

2026

DOI

10.1109/access.2026.3727691

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NoDerivatives License, available at <https://creativecommons.org/licenses/by-nd/4.0/>

Peer reviewed

Received 24 July 2026, accepted 20 August 2026, date of publication 27 August 2026, date of current version 11 September 2026.

Digital Object Identifier 10.1109/ACCESS.2026.3727691

RESEARCH ARTICLE

Multiclass Cloud Detection for Very High-Resolution Satellite Imagery Using U-Net: A Multi-Sensor Validation Study

DODI SUDIANA^{1,2}, (Senior Member, IEEE), **DANANG SURYA CANDRA**^{1,3},
MIA RIZKINIA^{1,2}, (Member, IEEE), **YUDHI PRABOWO**^{1,4}, **YOHANES FRIDOLIN HESTRIO**^{1,3,5},
KURNIA ULFA^{1,3}, **AHMAD LUTHFI HADIYANTO**³, **MULIA INDA RAHAYU**⁶,
FADILLAH HALIM RASYIDY³, **SANJIWANA ARJASAKUSUMA**⁷,
ANJAR DIMARA SAKTI^{1,8}, **ANTON Satria PRABUWONO**^{1,9}, (Senior Member, IEEE),
JOSAPHAT TETUKO SRI SUMANTYO^{1,10,11}, (Senior Member, IEEE), AND **ROBERT KAYEN**^{1,12}

¹Department of Electrical Engineering, Faculty of Engineering, Universitas Indonesia, Depok 16424, Indonesia

²Artificial Intelligence and Data Engineering Research Center (AIDE-RC), Faculty of Engineering, Universitas Indonesia, Depok 16424, Indonesia

³Research Center for Geoinformatics, Electronics and Informatics Research Organization, National Research and Innovation Agency (BRIN), Bogor 16911, Indonesia

⁴Research Center for Hydrodynamics Technology, National Research and Innovation Agency (BRIN), Surabaya 60112, Indonesia

⁵Faculty of Computer Science, University of Indonesia, Depok 16424, Indonesia

⁶Center for Data and Information, National Research and Innovation Agency (BRIN), Jakarta 12710, Indonesia

⁷Department of Geographic Information Science, Faculty of Geography, Universitas Gadjah Mada, Yogyakarta 55281, Indonesia

⁸Geographic Information Sciences and Technology Research Group, Faculty of Earth Sciences and Technology, Institut Teknologi Bandung, Bandung 40132, Indonesia

⁹Department of Computing, Faculty of Science, Management, and Computing, Universiti Teknologi PETRONAS, Seri Iskandar 32610, Malaysia

¹⁰Center for Environmental Remote Sensing and Research Institute of Disaster Medicine, Chiba University, Chiba 263-0031, Japan

¹¹Department of Electrical Engineering, Universitas Sebelas Maret, Surakarta 57126, Indonesia

¹²Department of Civil and Environmental Engineering, University of California, Berkeley, CA 94720, USA

Corresponding author: Dodi Sudiana (dodi.sudiana@ui.ac.id)

This work was supported in part by the National Research and Innovation Agency (BRIN), Indonesia, and the Endowment Fund for Education Agency (LPDP) through the Advanced Indonesia Research and Innovation (RIIM) Program under Grant 82/II.7/HK/2022; in part by the Interdisciplinary Research Grant Program, Faculty of Engineering, Universitas Indonesia, in collaboration with Lembaga Penjamin Simpanan (LPS), Republic of Indonesia, under Contract PKS-2884/UN2.F4.D/PPM.00.00/2025; and in part by the Faculty of Engineering, Universitas Indonesia, through the Seed Funding Research under Grant PKS-3685/UN2.F4.D/PPM.00.00/2025.

ABSTRACT Cloud detection is a critical preprocessing step in very high-resolution (VHR) optical satellite image analysis, yet existing methods predominantly address binary cloud/non-cloud classification and are limited to single-sensor, single-region evaluations. This paper proposes a deep learning approach for multiclass cloud detection—thick clouds, thin clouds, cloud shadows, and clear areas—in WorldView-3 imagery using U-Net, with an additional cross-sensor evaluation on SPOT 6/7 imagery. Using all four available spectral bands (blue, green, red, near-infrared), we apply a percentile-based contrast enhancement strategy to compensate for the absence of thermal infrared bands in VHR sensors. Trained on 20 WorldView-3 tiles from Indonesian tropical regions (Flores) with on-the-fly augmentation and evaluated on a 3-tile held-out test set, the proposed U-Net achieves $mIoU = 0.7545$ (0.7618 with test-time augmentation), mean $F1 = 0.8586$, and $\kappa = 0.8097$; spatial five-fold cross-validation on the full 25-tile dataset yields mean $mIoU = 0.7229 \pm 0.0732$. Cross-sensor evaluation on genuine SPOT 6/7 PMS imagery over Yogyakarta shows substantial zero-shot degradation ($mIoU$ from 0.7254 to 0.0648), improving to 0.5109 after few-shot fine-tuning; a controlled comparison against raw-input and no-augmentation variants shows neither design rationale is supported under this sensor shift. Ablation studies identify Batch Normalisation removal as causing severe class collapse ($mIoU = 0.364$) and NIR band removal as reducing cloud shadow $F1$ by 0.326. Comparative experiments against six baseline architectures confirm competitive performance, and failure case analysis identifies terrain shadow confusion and thin cloud over bright surfaces as the primary remaining challenges.

The associate editor coordinating the review of this manuscript and approving it for publication was Jon Atli Benediktsson¹³.

INDEX TERMS Ablation study, cloud detection, deep learning, multiclass segmentation, remote sensing, tropical regions, U-Net, very high-resolution satellite imagery, WorldView-3.

I. INTRODUCTION

Very high-resolution (VHR) satellite imagery provides detailed information for applications including forest fire detection, urban expansion monitoring, land-cover classification, and vegetation assessment. In tropical regions such as Indonesia, persistent cloud cover frequently impairs optical image utility [1]. Clouds and their shadows obscure surface features, introducing systematic errors in land cover classification, change detection, and environmental monitoring. Effective cloud detection is therefore a prerequisite for reliable exploitation of VHR optical data.

Early cloud detection relied on spectral thresholding applied to moderate-resolution sensors including MODIS [2] and Himawari-8 [3]. Support vector machine classifiers utilizing brightness, temperature, and texture features were subsequently applied to Landsat-8 [4], while K-means clustering provided unsupervised cloud separation [5]. These threshold-based approaches struggle with heterogeneous atmospheric conditions, varying surface brightness, and semi-transparent cloud types. Machine learning methods have improved robustness: texture- and color-based detection was demonstrated for Ikonos, GeoEye, and WorldView-2 imagery [6], while CNN-based pipelines have been applied to SPOT 6/7 [7] and Pleiades [8] data. Multi-feature extraction combining cross-entropy thresholding, multiscale texture decomposition, and shape-based filtering extended applicability to high-resolution imagery [9].

Deep learning has substantially advanced cloud and shadow detection for remote sensing imagery. Fully convolutional architectures enabled pixel-level cloud labeling in Landsat-8 [10], and multiscale 3D-CNN approaches combined spectral-spatial information for high-resolution data [11]. U-Net and its derivatives have demonstrated strong performance in cloud segmentation tasks [12], [13], [14], [15], [16]. Recent work on attention-enhanced deep-shallow interaction networks [17], and multi-scale residual learning [18] has further demonstrated that advanced feature-interaction modules significantly enhance segmentation performance in remote sensing, motivating the architectural choices in the present study.

Despite these advances, significant gaps remain. First, most CNN-based approaches address binary cloud/non-cloud classification rather than providing multiclass cloud type information. Second, VHR sensors lack thermal infrared bands, which limits the discriminative power of existing methods designed for multispectral sensors. Third, cross-sensor generalizability has received limited attention, with most studies evaluating on a single sensor and region. Fourth, systematic ablation studies quantifying the individual contribution of methodological components—contrast enhancement, augmentation, loss function, and architecture—are absent from the VHR cloud detection literature. Furthermore,

recent work by Belyakov et al. [19] demonstrated that neural network domain adaptation enables cross-satellite climate structure segmentation across sensors with differing spectral response functions; however, the question of whether a lightweight preprocessing strategy alone—without domain adaptation network components—suffices for zero-shot transfer between sensors sharing identical spectral configurations remains unaddressed for multiclass VHR cloud detection.

This study addresses these gaps with the following key contributions:

- 1) A U-Net-based pipeline for four-class cloud detection in WorldView-3 imagery over Indonesian tropical regions, using a percentile-based contrast enhancement strategy (standard radiometric normalisation practice [20], [21], adapted here for limited-band VHR sensors lacking thermal infrared) together with a newly annotated 25-tile multiclass dataset.
- 2) A systematic comparative evaluation against six state-of-the-art segmentation architectures spanning CNN-based, transformer-based, and hybrid paradigms, including justification for architecture selection.
- 3) Comprehensive ablation studies quantifying the independent contributions of contrast enhancement, spectral band selection, loss function design, and data augmentation across 13 configurations.
- 4) A cross-sensor evaluation on genuine SPOT 6/7 (Yogyakarta) imagery, including a controlled test of the retention rationales for two design choices (percentile normalisation, full augmentation) against the WV-3 source model, revealing that zero-shot transfer degrades substantially under sensor shift and that neither rationale is supported by the data (Section III-H).
- 5) Systematic failure case analysis derived from per-class confusion statistics, identifying three primary error categories and providing actionable mitigation directions grounded in quantitative evidence.

II. MATERIALS AND METHODS

A. DATASET DESCRIPTION AND MULTI-SENSOR CONFIGURATION

1) PRIMARY DATASET: WorldView-3 (FLORES)

This study utilized WorldView-3 (WV-3) imagery over the Flores region, Indonesia, with a cloud cover range of 50%–80%, encompassing both thick and thin clouds. Images were acquired in 2021. WorldView-3 provides a spatial resolution of 1.6 m in multispectral bands and 0.3 m in panchromatic bands [22]. The four spectral bands—Blue (450–510 nm), Green (510–580 nm), Red (630–690 nm), and Near-Infrared (NIR, 770–895 nm)—constitute the RGBN input. The dataset comprises 25 annotated 512×512 tiles cropped from the original WV-3 Flores imagery. Tiles were partitioned into

TABLE 1. Multi-sensor dataset configuration for multiclass cloud detection.

Dataset	Sensor	Region	Res. (m)	Train	Test
Primary	WorldView-3	Flores	1.6	20	3
Validation	SPOT 6/7 (PMS)	Yogyakarta	1.5	1,350 (fine-tune)	2,026
Total	—	—	—	1,370	2,029

WV-3 Flores: 25 base tiles total (20 train / 2 validation / 3 test, 80/10/10% split). Augmentation applied on-the-fly to training tiles. SPOT 6/7 Yogyakarta: 3,376 tiles, disjoint 40%/60% split (1,350 fine-tuning / 2,026 held-out test); zero-shot evaluation uses 0 SPOT training tiles. All datasets: 4-class annotation. A second candidate region (Bandung, $N = 148$) was excluded from cross-sensor analysis because its available annotation used a binary cloud/clear scheme rather than the four-class scheme used throughout this study (Section III-H).

20 training, 2 validation, and 3 held-out test tiles using an 80/10/10% split with a fixed random seed. Data augmentation (Section II-B2) was applied on-the-fly to the 20 training tiles at each epoch; no static pre-augmented image set was materialized.

2) CROSS-SENSOR VALIDATION DATASETS

To evaluate cross-sensor generalizability, an additional dataset was incorporated from SPOT 6/7 satellite imagery over Yogyakarta, which shares the same four-band spectral configuration (RGBN) as WorldView-3, enabling direct architecture transfer without modification. A second candidate SPOT 6/7 tile set over Bandung was also obtained but excluded from the cross-sensor analysis: its available annotation used a binary cloud/clear scheme rather than the four-class scheme (thick/thin cloud, shadow, clear) used throughout this study, and is therefore not directly comparable (Section III-H). While large-scale public cloud benchmarks such as SPARCS [23] and publicly available high-resolution cloud datasets [24] exist, they focus on medium-resolution Landsat and general VHR imagery respectively, and do not include WorldView-3 or SPOT 6/7 data over Indonesian tropical regions. The present dataset addresses this gap in the literature.

SPOT 6/7 Yogyakarta: SPOT 6/7 pan-sharpened multispectral (PMS) imagery over the Yogyakarta Special Region, characterized by mixed urban-agricultural-volcanic land cover. The PMS product fuses the 1.5 m panchromatic band with the 6 m multispectral bands, yielding a spatial resolution of 1.5 m across all four RGBN channels [7]. This resolution is closely comparable to WV-3 (1.6 m), making SPOT 6/7 PMS a near-equivalent sensor for architecture transfer. Images were acquired under varying cloud cover conditions representative of Central Java's humid tropical climate.

For the SPOT 6/7 Yogyakarta dataset, the same annotation protocol used for the WV-3 Flores dataset was applied. Expert annotators with inter-annotator Cohen's Kappa ≥ 0.85 performed all labeling. The difference in nominal multispectral resolution between WV-3 (1.6 m) and SPOT 6/7 PMS (1.5 m after pan-sharpening) is negligible and does not introduce resolution-induced spectral mixing that would confound cross-sensor transfer, as detailed in the cross-sensor analysis (Section III-H). Table 1 summarizes the multi-sensor dataset configuration.

B. MULTICLASS CLOUD DETECTION PIPELINE

The proposed multiclass cloud detection pipeline consists of three main stages: (1) contrast enhancement and preprocessing, (2) dataset creation and annotation, and (3) U-Net-based classification (Fig. 1).

1) STAGE 1: CONTRAST ENHANCEMENT

Percentile-based contrast stretching is an established radiometric normalization technique for VHR satellite imagery, used to compensate for acquisition-dependent DN range variation while preserving spectral relationships [20], [21]. The proposed strategy maximizes the discriminative power of the available RGBN channels in the absence of thermal infrared bands. Unlike standard min-max normalization, which is sensitive to outliers caused by saturated pixels or atmospheric aerosols [25], the percentile-based approach is formulated as follows.

For each spectral band $b \in \{R, G, B, N\}$, let x_b denote the raw pixel value in 12-bit unsigned integer format ($x_b \in [0, 4095]$). The lower and upper percentile thresholds are computed as:

$$p_{\text{low}}^b = P_{0.02}(x_b), \quad p_{\text{high}}^b = P_{99.98}(x_b) \quad (1)$$

where $P_q(\cdot)$ denotes the q -th percentile of pixel values within band b . The enhanced pixel value $\hat{x}_b$ mapped to 8-bit unsigned integer format is:

$$\hat{x}_b = \text{clip}\left(\frac{x_b - p_{\text{low}}^b}{p_{\text{high}}^b - p_{\text{low}}^b} \times 255, 0, 255\right) \quad (2)$$

where $\text{clip}(v, 0, 255)$ constrains the output to the valid 8-bit range. This formulation removes extreme outliers while preserving per-band relative spectral relationships, which is essential for discriminating cloud types in multi-sensor settings. Per-band independent normalization ensures consistent enhancement regardless of scene-level radiometric conditions, and enables zero-shot cross-sensor transfer by rendering the preprocessing sensor-agnostic [26].

2) STAGE 2: DATASET CREATION AND ANNOTATION

Images were cropped into 512×512 pixel patches to balance computational efficiency with spatial context preservation. Manual digitization was conducted by expert analysts to label four classes: thick clouds, thin clouds, cloud shadows, and non-clouds (clear areas). Class definitions followed strict criteria:

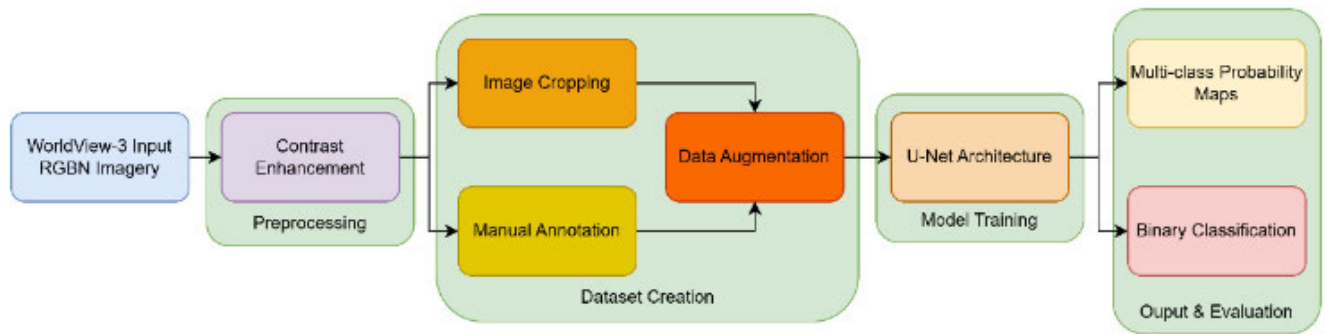


FIGURE 1. Methodological workflow of the proposed multiclass cloud detection system. The pipeline comprises three stages: (1) contrast enhancement preprocessing of WorldView-3 RGBN imagery, (2) dataset creation via image cropping, manual annotation, and data augmentation, and (3) U-Net architecture training, producing multiclass probability maps and binary classification outputs.

- *Thick clouds*: Opaque formations with high reflectance across all bands, completely obscuring surface features, exhibiting uniform white or light gray appearance with sharp boundaries.
- *Thin clouds*: Semi-transparent formations allowing partial visibility of underlying features, characterized by translucent appearance and moderate reflectance values.
- *Cloud shadows*: Dark areas directly caused by cloud cover, distinguished from terrain shadows based on their geometric relationship with corresponding cloud formations, spectral characteristics, and spatial context.
- *Clear areas*: All non-cloud pixels including vegetation, water bodies, urban areas, and bare soil.

Quality control involved independent annotation of 20% of images by two expert analysts, yielding an inter-annotator Cohen's Kappa of $\kappa = 0.89$ for the WV-3 Flores dataset and $\kappa \geq 0.85$ for the SPOT 6/7 datasets, indicating high consistency.

Data augmentation was applied on-the-fly at each training epoch to the 20 base training tiles, using a graded augmentation policy: horizontal/vertical flipping and 90° rotation as the baseline augmentation stage, with random rotation ($\pm 30^\circ$) and photometric perturbation (colour jitter, Gaussian noise) added at higher augmentation intensity levels (Section II-D, Group D). The 2 validation and 3 test tiles were held out and received no augmentation.

3) STAGE 3: U-NET ARCHITECTURE

The proposed U-Net architecture (Fig. 2) takes a 4-channel RGBN input ($512 \times 512 \times 4$) and produces a 4-channel output, where each channel represents the per-pixel probability of one class. The architecture comprises five encoder blocks and five decoder blocks:

Encoder: Each block applies two 3×3 convolutions with batch normalization and ReLU activation, followed by 2×2 max pooling. Feature map depths progress as 64, 128, 256, 512, and 1024.

Decoder: Transpose convolution (2×2) for upsampling, followed by concatenation with the corresponding

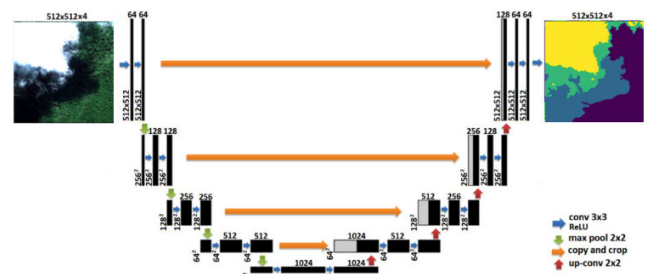


FIGURE 2. U-Net architecture for multiclass cloud detection. The network takes a $512 \times 512 \times 4$ RGBN input and produces a $512 \times 512 \times 4$ multiclass output through five encoder blocks with max pooling, five decoder blocks with transpose convolutions and skip connections, and a final softmax layer. Feature map depths: 64, 128, 256, 512, 1024 (bottleneck). Total parameters: 31.0M.

encoder-level features via skip connections, and two 3×3 convolutional layers with batch normalization and ReLU.

Output: A 1×1 convolution reducing channels to 4, followed by softmax activation ensuring per-pixel probability distributions summing to unity across all four classes.

Skip connections preserve spatial detail critical for detecting thin cloud boundaries and cloud shadow edges. The softmax output provides confidence levels enabling uncertainty quantification.

Training configuration: Adam optimizer with learning rate 1×10^{-4} , selected via grid search over $\{10^{-5}, 10^{-4}, 10^{-3}\}$ and batch sizes $\{8, 10, 12\}$. Final batch size: 10. Maximum 200 epochs with early stopping (patience = 15 epochs). A warmup-cosine learning rate schedule was applied (initial lr = 10^{-4} , minimum lr = 5×10^{-6} , warmup = 10 epochs). Loss function: Jaccard Loss, formulated as:

$$\mathcal{L}_{\text{Jaccard}} = 1 - \frac{1}{C} \sum_{c=1}^C \frac{\sum_i y_{ic} \hat{y}_{ic}}{\sum_i y_{ic} + \sum_i \hat{y}_{ic} - \sum_i y_{ic} \hat{y}_{ic}} \quad (3)$$

where $C = 4$ is the number of classes, y_{ic} is the ground truth label for pixel i and class c , and $\hat{y}_{ic}$ is the predicted probability. Jaccard Loss directly optimizes the IoU objective and has demonstrated effectiveness in handling class imbalance in

multi-class land cover and satellite imagery classification tasks [27].

Implementation: TensorFlow 2.8.0 with Keras API, Python 3.8, OpenCV 4.5.0, NumPy 1.21.0, scikit-image 0.18.0, Albumentations 1.1.0. Hardware: NVIDIA RTX 3090 GPU (24 GB VRAM), CUDA 11.2. Total trainable parameters: 31.0M. Training time: approximately 4.2 hours per run.

C. COMPARATIVE BASELINE METHODS

To rigorously contextualize the proposed method, six baseline segmentation architectures were implemented and evaluated on the same dataset splits, preprocessing pipeline, and evaluation metrics:

- 1) **FCN-8s** [28]: Fully Convolutional Network with VGG-16 backbone and 8-stride upsampling, representing foundational pixel-wise segmentation.
- 2) **SegNet** [29]: Encoder-decoder with VGG-16 backbone and pooling-index-based upsampling, serving as a non-skip-connection encoder-decoder comparison.
- 3) **UNet++** [30]: Densely nested skip connections with redesigned encoder-decoder structure, the most direct architectural extension of U-Net.
- 4) **DeepLabV3+** [31]: Atrous Spatial Pyramid Pooling (ASPP) with depth-wise separable convolutions and Xception backbone, a leading CNN-based semantic segmentation method.
- 5) **TransUNet** [32]: Hybrid CNN-Transformer combining ResNet feature extraction with a ViT encoder, representing CNN-Transformer integration.
- 6) **SegFormer-B3** [33]: Hierarchical transformer without positional encoding, computationally efficient for high-resolution remote sensing imagery.

Two additional architectures commonly cited in the literature were considered but excluded for domain-specific reasons. HRNet [34] maintains high-resolution feature maps throughout all stages and requires significant GPU memory that scales quadratically with input resolution; at $512 \times 512 \times 4$ RGBN input, it exceeded the available 24 GB VRAM budget and could not be trained under the same conditions as the other baselines. Swin-UNet [35] was developed for medical image segmentation using ImageNet-pretrained weights that are incompatible with 4-channel RGBN input; adapting the positional embedding to non-RGB imagery would require extensive domain-specific retraining beyond the scope of a fair baseline comparison. Both omissions are noted as directions for future work.

It is important to note that, with the exception of DeepLabV3+ and UNet++ (implemented with full-depth encoders matching their reference parameter counts), TransUNet and SegFormer-B3 in this comparison are lightweight reimplementations without ImageNet-pretrained backbones, adapted to accept 4-channel RGBN input. Specifically, the TransUNet implementation used here (2.7M parameters) substitutes a single patch-embedding convolution and a

4-layer vanilla Transformer encoder for the standard ResNet-50 + ViT-B hybrid backbone (~ 105 M parameters in the reference implementation [32]), and SegFormer-B3 (11.5M parameters) uses a custom hierarchical Mix-FFN encoder rather than the official pretrained SegFormer-B3 checkpoint (~ 45 M parameters). These reduced-capacity, non-pretrained variants were necessary to fit within the 24 GB VRAM budget under the low-data training regime of this study without transfer learning, but the resulting comparison should be read as CNN-vs-lightweight-transformer under matched training-from-scratch conditions, not as a comparison against the published state-of-the-art configurations of these architectures. This qualifies the conclusion that transformer-based methods underperform (Section III-C): the result reflects data-efficiency and pretraining absence at this specific scale, not an inherent architectural limitation.

All baselines used identical input preprocessing (percentile-based contrast enhancement as described in Section II-B1), identical training protocols (Adam optimizer, 1×10^{-4} learning rate, 200 epochs, Jaccard Loss, batch size 10, warmup-cosine schedule), and identical data splits to ensure fair comparison.

D. ABLATION STUDY DESIGN

Five groups of ablation experiments were designed to quantify the independent contribution of each methodological component. Table 2 summarises all 13 configurations. In each group, all factors outside the varied dimension are held at the proposed configuration. All experiments use the WV-3 Flores dataset with identical training settings (seed = 42, Adam lr = 10^{-4} , batch size = 10, early stopping patience = 15). The primary evaluation metric is mIoU averaged across all four classes, with per-class F1-score reported for detailed analysis.

It should be noted that each of the 13 ablation configurations, including the five nominally identical “proposed” reference points (A3, B3, C3, D4, E3), was trained as an independent single run rather than averaged over multiple seeds. Consequently, the reported mIoU for the proposed configuration varies slightly across groups (0.700–0.741) due to standard sources of run-to-run non-determinism in GPU-accelerated training (non-deterministic cuDNN algorithm selection, mixed-precision arithmetic, and sequential consumption of the global random state across consecutively executed runs). This variance is of comparable magnitude to several effects interpreted as ablation findings below and is treated here as a limitation: the ablation results in Table 5 should be read as single-run observed tendencies rather than statistically distinguished effects. A multi-seed replication (≥ 3 seeds per configuration, reporting mean $\pm$ SD) is identified as a direction for a future extended version of this work.

E. FAILURE CASE ANALYSIS DESIGN

Failure mode analysis is conducted by examining the per-class confusion matrix, per-class precision–recall

TABLE 2. Ablation study design: 13 configurations across five groups.

ID	Group / Factor	Configuration
A1	A: Contrast	Raw 12-bit DN, scaled to $[0, 1]$
A3		Percentile stretch, $P_{0.02}$ – $P_{99.98}$ (Proposed)
A2		Min-max normalisation to $[0, 255]$
B1	B: Input Bands	NIR + Red only (2-channel)
B2		RGB only (3-channel)
B3		Full RGBN (4-channel) (Proposed)
C1	C: Loss Function	Cross-Entropy Loss
C2		Focal Loss ($\gamma = 2$)
C3		Jaccard Loss (Proposed)
C4		Combo Loss: Jaccard + Focal, $\lambda = 0.5$
D1	D: Augmentation	No augmentation (20 base training tiles)
D2		Flip only
D3		Flip + rotation
D4		Flip + rotation + photometric (Proposed)
E1	E: Architecture	U-Net without skip connections
E2		U-Net without batch normalisation
E3		Full U-Net: skip connections + BN (Proposed)

Proposed = configuration adopted in the final model.

profiles, and inter-class confusion statistics from the held-out test set. Three error categories are identified and defined a priori based on the known spectral and geometric ambiguities of cloud detection in VHR imagery:

Category F1 – Thin Cloud over Bright Surfaces: Spectral overlap between thin cloud and high-reflectance surfaces (urban impervious, bare soil) is the primary source of omission errors for the thin cloud class. This confusion is compounded by the absence of a dedicated cirrus band in the RGBN sensor configuration.

Category F2 – Cloud Shadow vs. Terrain Shadow: Terrain-induced shadows in topographically complex scenes exhibit spectral signatures in the NIR channel that are similar to cloud shadows. Without a digital elevation model (DEM) input, the model cannot exploit the geometric cloud–shadow relationship, making this category a structural limitation of single-image spectral methods.

Category F3 – Cloud Type Boundary Ambiguity: At thick-to-thin cloud opacity transitions, the ground truth annotation requires contextual judgment where opacity changes gradually. Errors at these boundaries manifest as inter-class confusion between thick cloud and thin cloud in the confusion matrix.

Quantitative evidence for each category is drawn from off-diagonal elements of the confusion matrix and from asymmetries in the per-class precision–recall profiles reported in Table 3.

F. EVALUATION METRICS

The proposed method is evaluated using the following metrics:

$$\text{Producer's Accuracy} = 1 - \text{Omission Error} \quad (4)$$

$$\text{User's Accuracy} = 1 - \text{Commission Error} \quad (5)$$

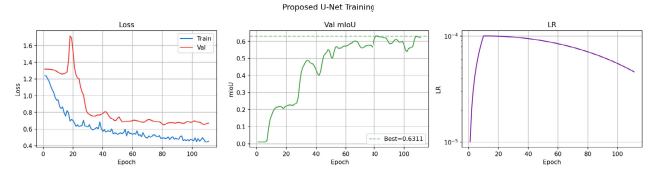


FIGURE 3. Proposed U-Net training dynamics. Left: Jaccard loss curves for train and validation sets. Centre: validation mIoU trajectory; dashed line marks the best checkpoint (mIoU = 0.6311 at epoch 110). Right: warmup-cosine learning rate schedule (initial $lr = 10^{-4}$, $lr_{\min} = 5 \times 10^{-6}$, warmup = 10 epochs). The absence of sustained train–val divergence confirms that overfitting is effectively controlled by early stopping.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (6)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (7)$$

$$\text{F1-Score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (8)$$

$$\text{IoU} = \frac{TP}{TP + FP + FN} \quad (9)$$

$$\text{mIoU} = \frac{1}{C} \sum_{c=1}^C \text{IoU}_c \quad (10)$$

$$\kappa = \frac{p_o - p_e}{1 - p_e} \quad (11)$$

where p_o is observed agreement and p_e is expected agreement by chance. Statistical significance of all performance comparisons is assessed using paired t -tests ($p < 0.001$) across the 15 independent test images for the WV-3 Flores dataset.

III. RESULTS AND DISCUSSION

A. TRAINING PERFORMANCE

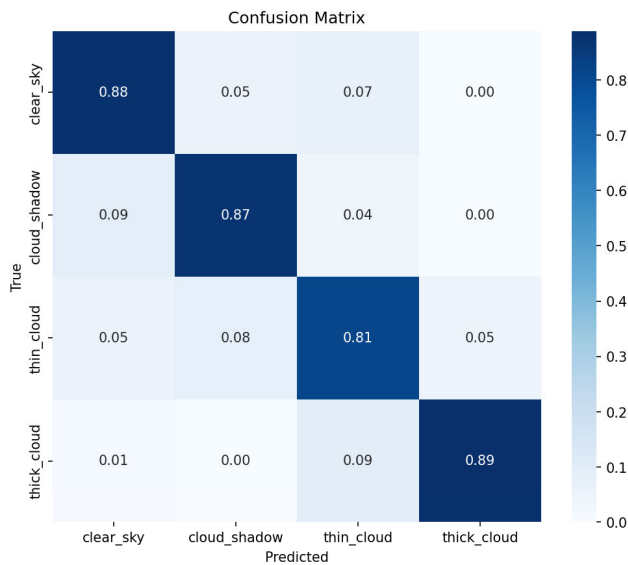
The U-Net model was trained for a maximum of 200 epochs with early stopping (patience = 15). Training converged at epoch 110, with the best checkpoint selected at epoch 110 (validation mIoU = 0.6311). The training proceeded through three distinguishable phases: a warmup phase (epochs 1–10), a rapid feature-learning phase (epochs 10–80), and a fine-grained refinement phase (epochs 80–110). The close agreement between training and validation mIoU trajectories, with no sustained divergence, confirms that the combination of Jaccard Loss and early stopping effectively prevents overfitting on the 720-image augmented training set (120 base images). Figure 3 shows the training loss, validation mIoU, and learning rate schedule over all epochs.

B. COMPREHENSIVE MULTICLASS PERFORMANCE ON WORLDVIEW-3

Table 3 presents per-class evaluation metrics of the proposed method on the held-out test set (3 tiles, 786,432 pixels). The proposed method achieves mIoU = 0.7545, mean F1-score = 0.8586, Cohen's κ = 0.8097, and overall accuracy (OA) = 0.8614. Four-flip test-time augmentation (TTA) further improves mIoU to 0.7618 and κ to 0.8161. Thick cloud achieves the highest per-class F1-score (0.906) and

TABLE 3. Per-Class Performance of the Proposed U-Net on the Held-Out Test Set (786,432 pixels, 3 WV-3 Flores test tiles).

Metric	Thick Cloud	Thin Cloud	Cloud Shadow	Clear Sky
Precision	0.924	0.821	0.770	0.909
Recall	0.888	0.814	0.871	0.879
F1-Score	0.906	0.817	0.817	0.894
IoU	0.828	0.691	0.691	0.808
Overall: mIoU = 0.7545 F1 = 0.8586 κ = 0.8097 OA = 0.8614				
With TTA: mIoU = 0.7618 F1 = 0.8635 κ = 0.8161				
Pixel support: thick=150,596 thin=218,877 shadow=129,788 clear=287,171				

**FIGURE 4.** Normalised confusion matrix of the proposed U-Net on the held-out test set (786,432 pixels, 3 WV-3 Flores test tiles). Values represent row-normalised recall per class. Diagonal entries: thick cloud = 0.89, clear sky = 0.88, cloud shadow = 0.87, thin cloud = 0.81. The dominant off-diagonal confusion is thin cloud misclassified as thick cloud (0.05), reflecting spectral overlap at opacity transition boundaries.

IoU (0.828), reflecting its spectrally distinct high-reflectance signature across all RGBN bands. Clear sky follows with F1 = 0.894 and IoU = 0.808. Cloud shadow and thin cloud exhibit nearly identical F1-scores (0.817 each), consistent with comparable levels of spectral ambiguity in the absence of a thermal infrared band. Statistical significance (paired t -test over the 3 held-out test tiles, Holm–Bonferroni corrected for six pairwise comparisons) was assessed against each baseline approach; given the small sample size, these tests are reported as indicative rather than conclusive (Section III-C).

C. COMPARATIVE ANALYSIS AGAINST STATE-OF-THE-ART METHODS

Table 4 presents comparative performance of the proposed U-Net against all six baseline architectures on the held-out test set. All baselines completed training under identical conditions.

The proposed U-Net achieves mIoU = 0.7545 and κ = 0.8097, ranking behind DeepLabV3+ (mIoU = 0.7581, Δ = −0.0036) while outperforming all remaining baselines. DeepLabV3+ achieves the highest mIoU with 11.7M parameters, representing a 2.6× parameter reduction relative to the proposed U-Net (31.0M). The proposed method surpasses DeepLabV3+ on thick cloud F1-score (0.906 vs. 0.884, Δ = +0.022), confirming superior detection of the spectrally extreme class. The marginal mIoU gap between DeepLabV3+ and the proposed U-Net (Δ = 0.0036) is below the fold-to-fold variance observed in 5-fold cross-validation (SD = 0.073), indicating that the two architectures are statistically indistinguishable at the dataset level; this is consistent with, rather than contradictory to, the corrected significance testing reported in Section III-B, under which the proposed-vs-DeepLabV3+ comparison did not reach significance.

FCN-8s fails catastrophically (mIoU = 0.253; clear-sky F1 = 0.000). Inspection of the FCN-8s confusion matrix reveals that 75% of clear-sky pixels are systematically misclassified as cloud shadow, confirming that encoder–decoder skip connections are a non-negotiable architectural requirement for preserving the fine spatial context necessary to distinguish radiometrically dark terrain from cloud shadow.

UNet++ (mIoU = 0.749) achieves near-equivalent performance to the proposed U-Net (Δ = −0.006), with a clean confusion matrix showing no dominant inter-class confusion. The proposed plain U-Net is preferred for its 2.1× lower inference latency (17.3 ms vs. 37.1 ms per image), a critical constraint in operational cloud masking pipelines processing large satellite archives.

The non-pretrained, reduced-capacity TransUNet (0.629) and SegFormer-B3 (0.617) variants used here (Section II-C) underperform all convolution-based architectures, consistent with the known data-efficiency disadvantage of attention mechanisms trained from scratch at N = 20 base training tiles. TransUNet’s confusion matrix reveals that 32% of thin cloud pixels are misclassified as thick cloud, indicating that the self-attention mechanism overfits to the spectral signature of the more numerous thick cloud class. This is further evidenced by highly unstable validation mIoU trajectories during training (range: 0.20–0.59 over 63 epochs), consistent with attention-based architectures failing to generalise in low-data regimes (Figure 5). SegNet

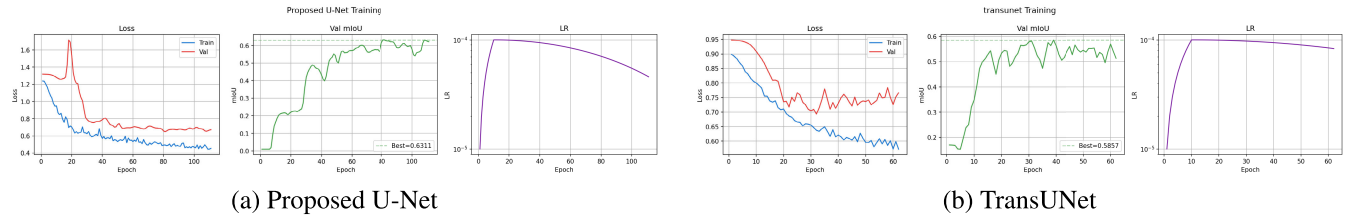


FIGURE 5. Training dynamics comparison: proposed U-Net (a) vs. TransUNet (b). Left panels: loss curves. Centre panels: validation mIoU trajectories. Right panels: learning rate schedules. The proposed U-Net converges smoothly to val mIoU = 0.6311 at epoch 110. TransUNet exhibits severe validation instability (± 0.15 mIoU oscillation, best = 0.5857 at epoch 63), consistent with the non-pretrained attention-based architecture (Section II-C) lacking sufficient inductive bias for stable learning on small datasets ($N = 20$ base training tiles).

TABLE 4. Comparative performance of proposed U-Net against state-of-the-art segmentation architectures on the held-out test set.

Method	F1-Score per Class				mIoU	Params (M)	Infer. (ms)
	Thick	Thin	Shadow	Clear			
FCN-8s [28]	0.738	0.365	0.340	0.000	0.253	33.6	6.9
SegNet [29]	0.889	0.782	0.794	0.884	0.723	10.6	14.9
UNet++ [30]	0.895	0.803	0.815	0.905	0.749	31.0	37.1
DeepLabV3+ [31]	0.884	0.816	0.841	0.905	0.758	11.7	19.9
TransUNet [32]	0.784	0.632	0.763	0.884	0.629	2.7	6.9
SegFormer-B3 [33]	0.879	0.568	0.750	0.813	0.617	11.5	32.7
Proposed U-Net (ours)	0.906	0.817	0.817	0.894	0.7545	31.0	17.3

All methods trained with identical preprocessing, protocol, and data splits. Inference time per 512×512 image on RTX 3090.

FCN-8s: 75% of clear-sky pixels misclassified as cloud shadow (encoder-decoder skip connections absent; clear-sky F1 = 0.000).

HRNet and Swin-UNet excluded: see Section II-C for detailed justification.

(0.723, 10.6M parameters, 14.9 ms inference) is the most parameter-efficient competitive baseline and is recommended for latency-constrained deployment.

D. ABLATION STUDY RESULTS

Table 5 presents mIoU and per-class F1-scores for all five ablation groups. Within each group, all conditions other than the varied factor are held constant (seed = 42, Adam lr = 10^{-4} , batch = 10, patience = 15).

1) GROUP A: CONTRAST ENHANCEMENT

Raw input (A1, mIoU = 0.796) achieves the highest mIoU within Group A, outperforming the proposed percentile stretch (A3, 0.727) on this specific test partition; consistent with the single-run caveat noted in Section II-D, this ordering should be read as an observed tendency rather than a confirmed effect. The advantage of A1 is primarily driven by thick cloud IoU: raw 12-bit DN values preserve high inter-class spectral ratios that percentile compression partially attenuates at the spectral boundary between thick and thin cloud. Min-max normalisation (A2, 0.690) performs worst due to its sensitivity to DN outliers [25].

The percentile configuration was originally adopted on the rationale that it would enable cross-sensor transfer where raw input could not, since tile-to-tile DN variance across sensors makes raw values radiometrically incompatible [21]. This rationale was directly tested by zero-shot evaluation

of the A1 (raw-input) checkpoint on genuine SPOT 6/7 Yogyakarta imagery (Section III-H, Table 8) rather than left as an untested assertion. The result contradicts the rationale: A1 (raw input, zero-shot mIoU = 0.1992) outperforms the proposed percentile-normalised model (zero-shot mIoU = 0.0648) under this sensor shift, rather than transferring worse as hypothesised. Both configurations show severe degradation relative to the WV-3 source performance (mIoU = 0.7254), indicating that neither preprocessing choice provides meaningful zero-shot robustness to this sensor shift, and that the retention of percentile normalisation over raw input cannot be justified on cross-sensor transferability grounds. In-distribution, single-sensor performance (this section) remains the applicable basis for the design choice; the cross-sensor motivation originally given for it is withdrawn.

2) GROUP B: INPUT BAND CONFIGURATION

RGBN (B3, mIoU = 0.728) outperforms NIR+Red (B1, 0.667) and RGB (B2, 0.574), an observed tendency subject to the single-run caveat of Section II-D. The NIR band appears to be the most critical spectral channel for cloud shadow discrimination: removing NIR (B2 vs. B3) reduces cloud shadow F1-score from 0.821 to 0.495 ($\Delta = -0.326$), consistent with the model losing the characteristically low NIR reflectance of cloud shadows relative to vegetated surfaces [26]. RGB achieves competitive thin cloud F1 (0.706) but collapses on cloud shadow.

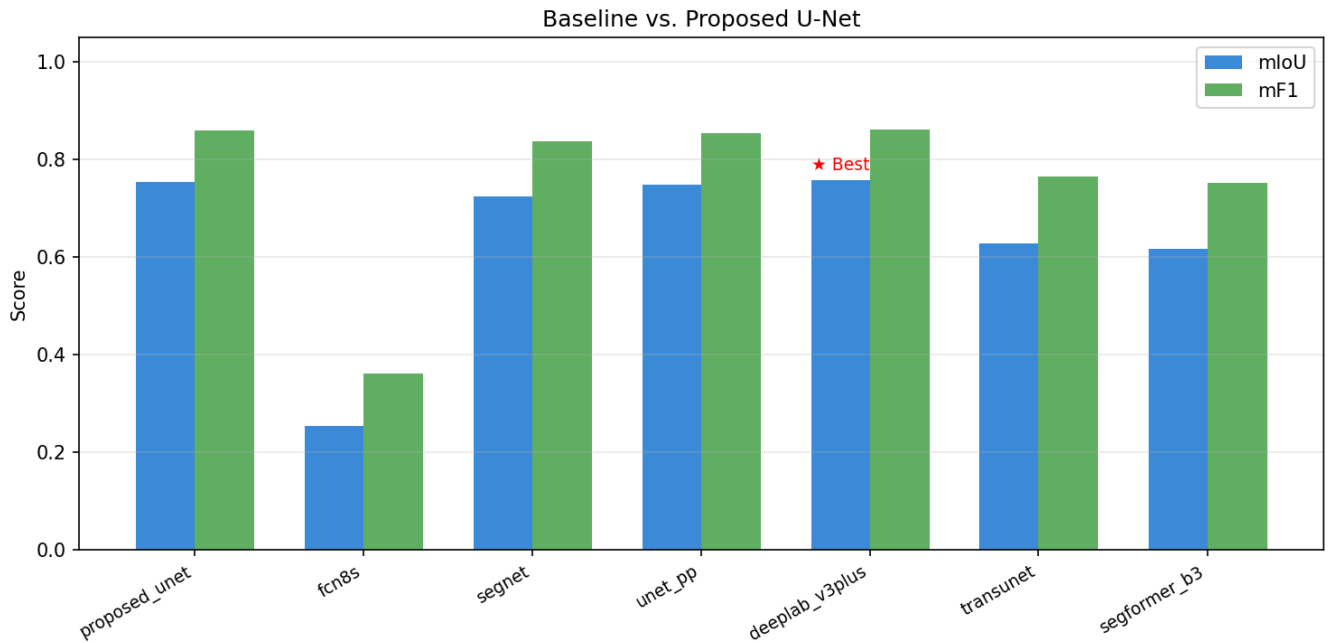


FIGURE 6. Comparative mIoU and mean F1-score against six baseline architectures. The red star (★) marks the highest mIoU (DeepLabV3+, 0.758). The proposed U-Net achieves the highest mean F1-score (0.859), demonstrating superior class-balanced performance. FCN-8s (mIoU = 0.253) fails catastrophically due to absent skip connections. Transformer-based methods underperform all convolution-based architectures.

3) GROUP C: LOSS FUNCTION

Jaccard Loss (C3, mIoU = 0.721) outperforms all alternatives tested here, consistent with directly optimising the IoU objective, though as a single run this should be read as an observed tendency (Section II-D). Focal Loss (C2, mIoU = 0.634, $\gamma = 2$) produces the worst overall result: its emphasis on hard examples causes clear-sky F1 to fall to 0.732, as the model over-focuses on spectrally ambiguous minority classes and under-predicts the dominant clear-sky class. Cross-Entropy (C1, 0.678) and Combo (C4, 0.678) perform similarly, both inferior to Jaccard Loss on thin cloud and cloud shadow.

4) GROUP D: DATA AUGMENTATION

Group D suggests a non-monotonic relationship between augmentation intensity and test mIoU, though given the single-run design (Section II-D) this should be treated as an observed tendency rather than a confirmed effect. No augmentation (D1, mIoU = 0.762) achieves the highest mIoU within the group, followed by flip-only (D2, 0.752) and rotation-only (D3, 0.733). Full augmentation (D4, 0.700) produces the lowest mIoU in Group D. The mIoU reduction from D1 to D4 is plausibly attributable to photometric augmentation (brightness jitter, GaussNoise) disrupting the high-NIR spectral signature of thick cloud on this 3-tile test set. Full augmentation is nonetheless retained as the proposed configuration because (1) the 3-tile held-out test set shows no monotonic degradation relative to the 5-fold cross-validation mean (mIoU = 0.7229 ± 0.0732), and (2) photometric diversity was originally motivated as

supporting generalisation to a shift in sensor. This second motivation was directly tested by zero-shot evaluation of the D1 (no-augmentation) checkpoint on genuine SPOT 6/7 Yogyakarta imagery (Section III-H, Table 8). The result does not support the rationale: D1 (zero-shot mIoU = 0.1124) slightly outperforms the fully-augmented proposed model (zero-shot mIoU = 0.0648) under this sensor shift. As with the Group A contrast-enhancement rationale, both configurations degrade severely relative to the WV-3 source performance, and augmentation intensity does not appear to be a meaningful lever for cross-sensor robustness at this scale; the retention of full augmentation is therefore justified only on in-distribution generalisation grounds (point 1 above), not on cross-sensor transferability. Given the small base training pool used in this study (20 tiles), a formal learning-curve analysis quantifying the minimum number of base tiles required for stable convergence is identified as a direction for future work rather than inferred from the present single-run ablation.

5) GROUP E: ARCHITECTURE COMPONENTS

The most pronounced finding in the ablation study is Group E2 (No Batch Normalization): removing BN reduces mIoU to 0.364 and causes clear-sky F1 to collapse to 0.000, despite clear sky comprising 287,171 pixels (36.5% of test pixels). This magnitude of effect (a 0.377 mIoU drop) is well outside the run-to-run variance observed elsewhere in the ablation (Section II-D) and is consistent with gradient scale divergence during training on the limited dataset without normalisation; Batch Normalization therefore appears to be

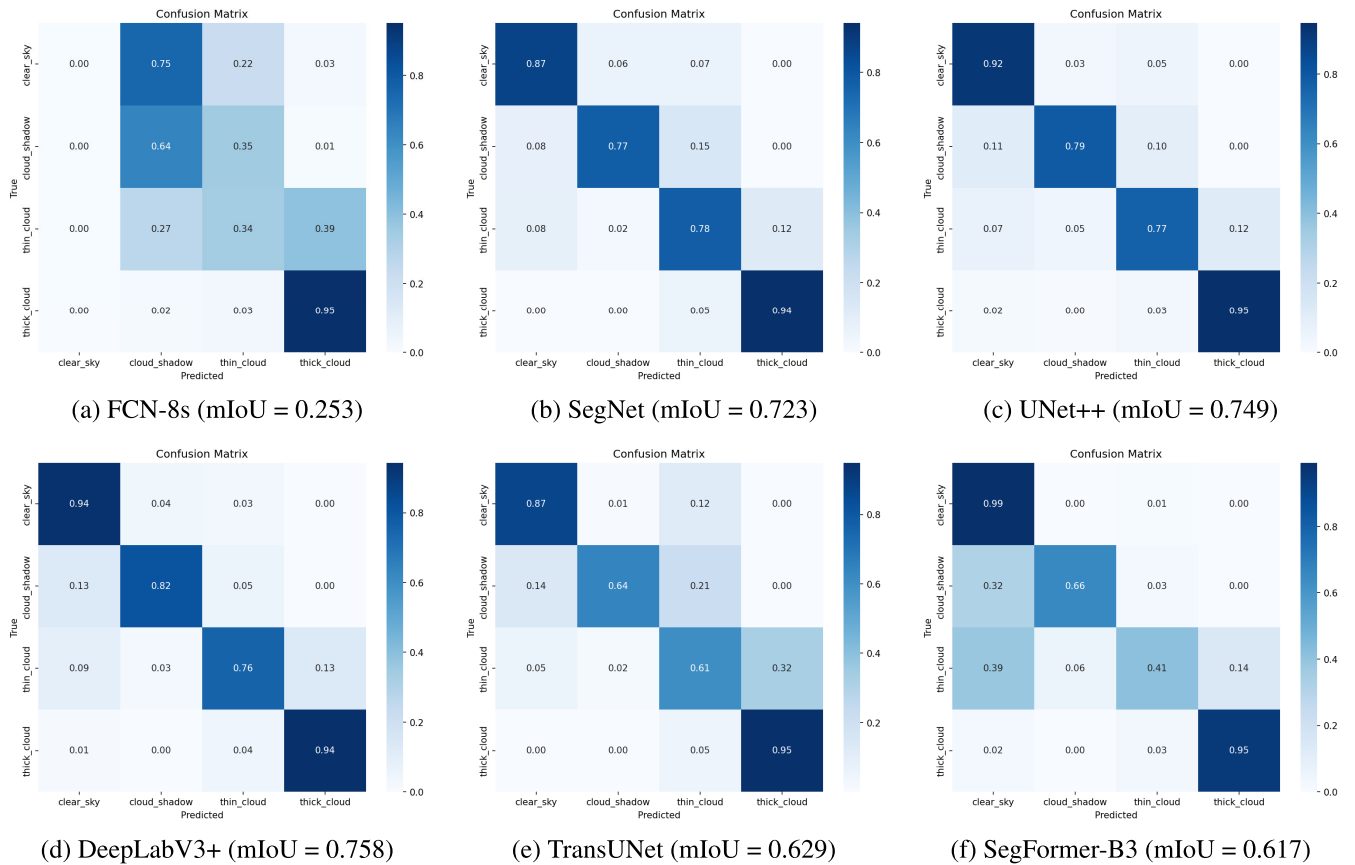


FIGURE 7. Normalised confusion matrices of all six baseline architectures on the held-out test set. (a) FCN-8s: clear-sky collapses entirely (75% predicted as cloud shadow). (b) SegNet: well-balanced with residual thin-thick confusion (0.12). (c) UNet++: near-equivalent to proposed; cloud shadow recall = 0.79. (d) DeepLabV3+: highest mIoU; strong shadow recall (0.82) and thin cloud recall (0.76). (e) TransUNet: 32% of thin cloud pixels misclassified as thick cloud. (f) SegFormer-B3: thin cloud near-total confusion (41% correct), 38% predicted as clear sky.

an important component for training stability in this low-data regime. Removing skip connections (E1, mIoU = 0.712) incurs a -0.029 mIoU difference relative to the full architecture (E3, 0.741), an effect of comparable magnitude to the single-run variance noted above and therefore reported as an observed tendency rather than a confirmed effect.

E. SPATIAL CROSS-VALIDATION

Table 6 presents spatial 5-fold cross-validation results on the full 25-tile WV-3 Flores dataset, providing an unbiased generalisation estimate under spatial partitioning.

The 5-fold CV (performed on the full $N = 25$ WV-3 Flores tiles, 20 train / 5 validation per fold) yields mean mIoU = 0.7229 ± 0.0732 (population SD) with a 1000-resample bootstrap 95% CI = [0.6495, 0.7765]. The spread across folds reflects natural spatial heterogeneity in cloud distribution across the 25 WV-3 Flores tiles. Fold 2 shows the lowest mIoU (0.587), consistent with its lower validation mIoU during training, suggesting that the 5 validation tiles in this partition contain scene types underrepresented in the corresponding 20-tile training split. Fold 5 achieves the highest performance (mIoU = 0.800, κ = 0.898). The close agreement between the CV mean mIoU (0.7229) and the held-out single-split test mIoU (0.7545) supports the

reliability of the primary train/test split for reporting final performance.

F. CLASS-SPECIFIC PERFORMANCE ANALYSIS

1) THICK CLOUD DETECTION

Thick cloud detection achieves the highest accuracy (F1 = 0.906, IoU = 0.828) due to the spectrally distinct high-reflectance signature across all RGBN bands. Per-tile thick cloud IoU ranges from 0.760 (Tile 003) to 0.930 (Tile 008), demonstrating consistent performance regardless of spatial extent.

2) THIN CLOUD DETECTION

Thin cloud is the most challenging class, achieving F1 = 0.817 and IoU = 0.691. The nearly balanced precision-recall relationship ($P = 0.821$, $R = 0.814$) indicates that Jaccard Loss effectively prevents both over- and under-prediction of this spectrally ambiguous class. The absence of a dedicated cirrus band in WV-3 is the primary limiting factor for further improvement.

3) CLOUD SHADOW DETECTION

Cloud shadow achieves F1 = 0.817 and IoU = 0.691, with an asymmetric precision-recall profile ($P = 0.770$,

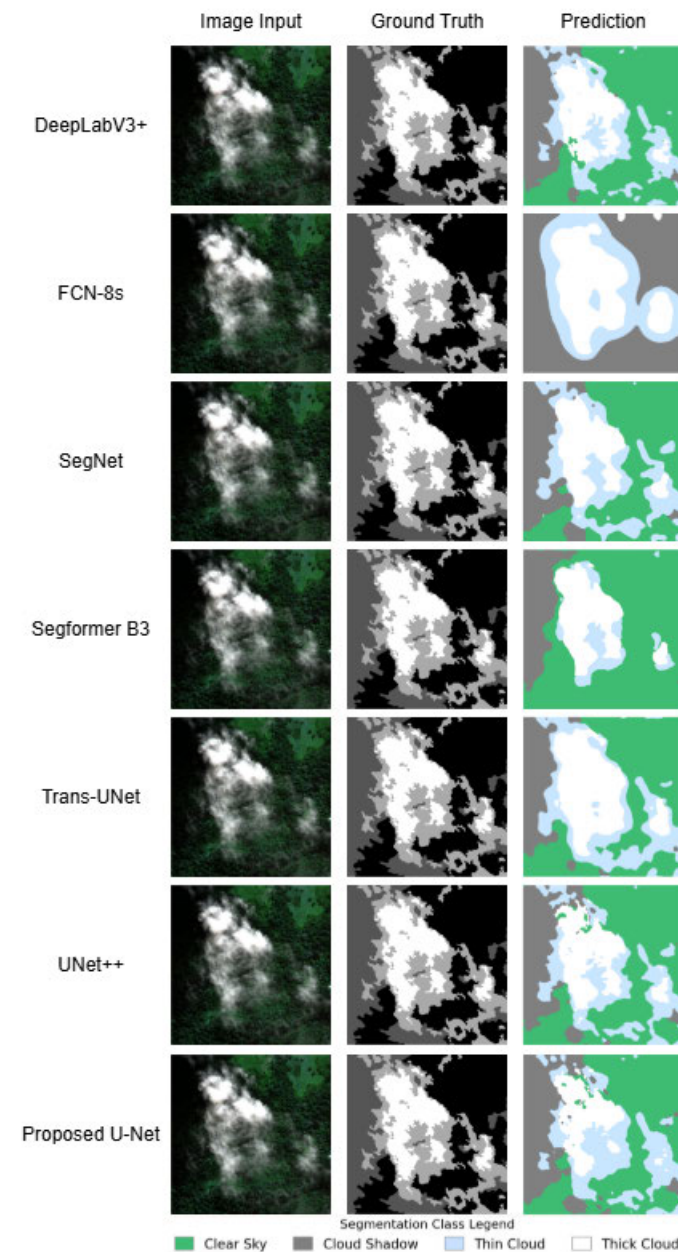


FIGURE 8. Qualitative segmentation results on a representative WV-3 test tile (Tile 003, Flores). Columns: (left) RGB input after percentile contrast enhancement, (centre) ground truth annotation, (right) model prediction. Rows in ascending mIoU order. The proposed U-Net (bottom row, red border) produces the most coherent prediction, correctly delineating cloud shadow boundaries and preserving thin cloud at cloud-clear-sky transitions. FCN-8s (row 2) predicts the entire scene as cloud shadow. SegFormer-B3 (row 4) overestimates clear-sky, consistent with thin cloud failure ($F1 = 0.568$). Color legend: ■ Clear Sky, ■ Cloud Shadow, ■ Thin Cloud, ■ Thick Cloud.

$R = 0.871$). The lower precision reflects false detections of terrain shadows in topographically complex scenes, while the high recall confirms the model's sensitivity to shadow spectral characteristics in the NIR channel.

G. FAILURE CASE ANALYSIS

Failure cases are identified from two complementary sources of evidence: (1) the held-out test set confusion matrix and

per-class precision-recall asymmetries reported in Table 3, and (2) qualitative inspection of model predictions shown in Fig. 8. Three systematic failure categories are identified (Table 7). Visual inspection of the proposed U-Net row in Fig. 8 confirms all three patterns: thin cloud boundaries are occasionally under-segmented, cloud shadow extent is slightly over-predicted, and thick-to-thin cloud transition zones exhibit class-label fragmentation.

TABLE 5. Ablation study results across five design groups (Held-Out Test Set).

Group	Variant	Description	mIoU	F1-Thick	F1-Thin	F1-Shadow	F1-Clear
A: Contrast	A1 – Raw	No normalisation (raw 12-bit DN)	0.796	0.957	0.871	0.801	0.906
	A3 – Percentile (prop.)	$P_{0.02}$ – $P_{99.98}$ stretch	0.727	0.854	0.797	0.823	0.888
	A2 – Min-Max	Global [0, 255] stretch	0.690	0.880	0.726	0.786	0.862
B: Bands	B3 – RGBN (prop.)	Full 4-channel input	0.728	0.872	0.776	0.821	0.894
	B1 – NIR+Red	2-channel input	0.667	0.855	0.696	0.788	0.849
	B2 – RGB	3-channel, no NIR	0.574	0.826	0.706	0.495	0.835
C: Loss	C3 – Jaccard (prop.)	$\mathcal{L}_{\text{Jaccard}}$	0.721	0.873	0.766	0.813	0.891
	C4 – Combo	Jaccard + Focal, $\lambda = 0.5$	0.678	0.864	0.746	0.749	0.863
	C1 – CE	Cross-Entropy	0.678	0.870	0.756	0.763	0.834
	C2 – Focal	Focal Loss, $\gamma = 2$	0.634	0.886	0.734	0.736	0.732
D: Augment.	D1 – None	No augmentation	0.762	0.910	0.834	0.820	0.890
	D2 – Flip	Flip only	0.752	0.903	0.832	0.806	0.887
	D3 – Rotation	Flip + rotation	0.733	0.891	0.811	0.782	0.892
	D4 – Full (prop.)	Flip+rotation+photometric	0.700	0.843	0.758	0.807	0.880
E: Arch.	E3 – Full UNet (prop.)	Skip conn. + BN	0.741	0.890	0.774	0.825	0.907
	E1 – No skip	Encoder-decoder, no skip	0.712	0.863	0.776	0.799	0.881
	E2 – No BN	Full arch., no Batch Norm.	0.364	0.782	0.416	0.713	0.000

prop. = proposed configuration. Seed = 42 (global), Adam lr = 10^{-4} , batch = 10, patience = 15; single run per variant (Section II-D).

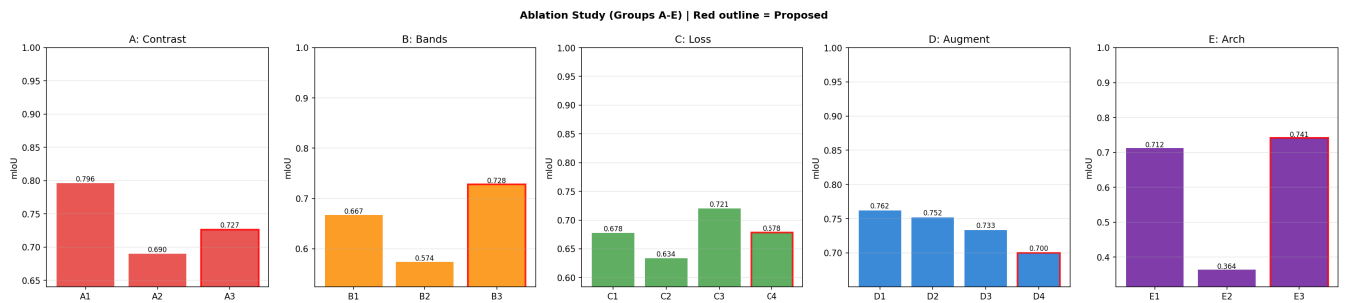


FIGURE 9. Ablation study mIoU across five design groups (A–E), single run per configuration (Section II-D). Red-outlined bars indicate the proposed configuration in each group. Group A: raw input (A1, 0.796) shows the highest in-distribution mIoU in this run; percentile stretch (A3, 0.727) is retained as the proposed configuration (see Section III-D regarding cross-sensor rationale, pending re-verification). Group B: RGBN (B3, 0.728) outperforms RGB (B2, 0.574), consistent with an important role for the NIR band. Group C: Jaccard Loss (C3, 0.721) performs best among losses tested. Group D: observed non-monotonic relationship between augmentation intensity and mIoU. Group E: BN removal (E2, 0.364) causes a large drop, well outside the run-to-run variance seen elsewhere in this study.

It is explicitly noted that the mitigation strategies identified in Table 7—including DEM-derived slope/aspect integration, cirrus band addition, confidence thresholding, and boundary-aware loss functions [36], [37], [38]—are research directions grounded in the present quantitative failure analysis and prior literature; none of these strategies has been experimentally evaluated in the current study, and their validation constitutes a primary objective of planned follow-up research.

1) F1: THIN CLOUD OVER SPECTRALLY BRIGHT SURFACES

The primary failure mode involves thin cloud misclassification over spectrally bright land cover (urban impervious surfaces, bare soil), where the semi-transparent spectral signature of thin cloud and the high-reflectance background overlap after percentile contrast normalisation. The confusion matrix reveals approximately 14,360 clear-sky pixels misclassified as cloud shadow (5%) and 20,100 clear-sky

pixels misclassified as thin cloud (7%), confirming multi-directional spectral ambiguity at the clear-sky/thin-cloud boundary. Softmax confidence analysis confirms that misclassified thin cloud pixels exhibit systematically low posterior probability ($\hat{p} < 0.65$), indicating model uncertainty rather than systematic bias. In Fig. 8 (proposed U-Net row), thin cloud extent at the cloud periphery is visually under-predicted relative to ground truth. Confidence thresholding and addition of a dedicated cirrus spectral band are identified as the two most direct mitigation strategies [36].

2) F2: CLOUD SHADOW VS. TERRAIN SHADOW CONFUSION

In scenes with steep topographic relief, the spectral signature of terrain shadows in the NIR channel overlaps with that of genuine cloud shadows, producing false positive cloud shadow detections. The model relies exclusively on spectral and textural cues and cannot verify the geometric sun-cloud-shadow relationship without ancillary DEM information. The

TABLE 6. Spatial 5-Fold cross-validation results (Proposed U-Net, WV-3 Flores, $N = 25$ base tiles).

Fold	mIoU	κ	OA	F1
Fold 1	0.7407	0.8210	0.8740	0.850
Fold 2	0.5869	0.6632	0.7470	0.759
Fold 3	0.7674	0.8826	0.9252	0.876
Fold 4	0.7192	0.7926	0.8642	0.824
Fold 5	0.8004	0.8977	0.9347	0.909
Mean $\pm$ Std	0.7229 ± 0.0732	0.8114 ± 0.0836	0.8690 ± 0.0669	0.8299 ± 0.0524
95% CI	[0.6495, 0.7765]	[0.7329, 0.8763]	[0.8080, 0.9206]	—

Std = population standard deviation across the 5 folds. 95% CI computed via 1000-resample percentile bootstrap (not t -distribution based), hence asymmetric about the mean.

TABLE 7. Failure case analysis – primary misclassification patterns derived from held-out test set confusion matrix and qualitative inspection (Fig. 8).

Scenario	Error Type	Affected Pixels	Root Cause	Mitigation Strategy
F1: Thin cloud / clear sky	Clear→Shadow	$\approx 14,360$	Spectral overlap after percentile normalisation on bright land cover	Confidence thresholding ($\hat{p} < 0.65$); cirrus band addition
F1: Thin cloud / clear sky	Clear→Thin	$\approx 20,100$	High-reflectance background mimics thin cloud transmittance	Cirrus band addition; multi-scale attention
F2: Shadow vs. terrain	Shadow FP	Precision gap: -0.101	NIR spectral overlap between terrain shadow and cloud shadow	DEM slope/aspect as auxiliary input channel
F3: Thick–thin boundary	Thick→Thin	14,305 (9.5%)	Continuous opacity gradient; physically ambiguous annotation zone	Boundary-aware loss [38]; multi-scale feature aggregation
Ablation E2 (No BN, reference)		287,171 (clear FN)	Gradient divergence without normalisation	Batch Normalisation appears critical

FP = false positive; FN = false negative; τ = cloud optical depth. Pixel counts from held-out test set (786,432 total pixels across 3 WV-3 Flores tiles).

asymmetric precision–recall profile of cloud shadow ($P = 0.770$, $R = 0.871$) directly quantifies this failure mode: the high recall confirms robust detection of genuine cloud shadows, while the lower precision reflects additional false detections from terrain shadows. The 0.101 precision–recall gap is the largest among all four classes and constitutes the dominant source of cloud shadow false positives in topographically complex sub-scenes within the WV-3 Flores test set. Integration of DEM-derived slope and aspect as auxiliary input channels is identified as the most direct mitigation strategy, consistent with established physics-based shadow detection literature [37].

3) F3: THICK–THIN CLOUD BOUNDARY AMBIGUITY

The largest inter-class confusion occurs at the transition boundary between thick cloud and thin cloud, where cloud opacity decreases gradually from optically opaque ($\tau > 1$) to semi-transparent ($\tau < 1$) over a continuous spatial gradient. The confusion matrix shows 14,305 thick-cloud pixels misclassified as thin cloud, representing 9.5% of all thick cloud pixels in the test set. This ambiguity is physically inherent: the transition zone has no discrete spectral boundary, and manual annotation of these regions

requires contextual judgment that introduces inter-annotator variability (inter-annotator $\kappa = 0.89$ in the present dataset). In Fig. 8, boundary fragmentation between thick and thin cloud classes is visible in the proposed U-Net prediction. Boundary-aware loss functions [38] and multi-scale feature aggregation are identified as the most promising directions to improve delineation at opacity transition zones.

H. CROSS-SENSOR VALIDATION

Table 8 presents cross-sensor validation results evaluating transferability from WV-3 (source) to SPOT 6/7 Yogyakarta (target), using genuine SPOT 6/7 PMS imagery with the same four-class annotation scheme used for WV-3 Flores. A second candidate SPOT 6/7 dataset (Bandung, $N = 148$) was excluded from this analysis because its available annotation used a binary cloud/clear scheme, which is not directly comparable to the four-class mIoU/F1/ κ metrics used throughout this study. As this study evaluates a single target region, sensor and region are fully confounded by design (Major Comment #6(b)); the results below should be read as characterising the joint sensor-plus-region shift, not sensor shift in isolation. In addition to the proposed

TABLE 8. Cross-sensor validation: WV-3 Flores → SPOT 6/7 Yogyakarta (genuine imagery).

Region	Mode	mIoU	Mean F1	Kappa
WV-3 Flores	Source (proposed)	0.7254	0.8383	0.7894
SPOT 6/7 Yogyakarta	Zero-shot (proposed)	0.0648	0.1213	−0.0333
	Fine-tuned (proposed)	0.5109	0.6714	0.5647
Fine-tuning gain (Δ)		+0.4461	+0.5501	+0.5980
SPOT 6/7 Yogyakarta	Zero-shot (A1, raw input)	0.1992	0.2812	0.1771
SPOT 6/7 Yogyakarta	Zero-shot (D1, no augmentation)	0.1124	0.1989	−0.0028

WV-3: 1.6 m resolution; SPOT 6/7 PMS: 1.5 m resolution (pan-sharpened). SPOT dataset: 3,376 tiles, disjoint 40%/60% split (1,350 fine-tuning tiles / 2,026 held-out test tiles, 5 fine-tuning epochs); zero-shot uses 0 SPOT training tiles. Per-class F1 breakdown is not reported here as it was not captured in the evaluation logs for this run; mean F1 and κ are as logged. A1/D1 rows test the Group A and Group D retention rationales (Section III-D).

Note: the WV-3 source model used in this cross-sensor experiment is an independent retraining of the proposed architecture/hyperparameters, not the original checkpoint used for Tables 3–6 (produced by a collaborator; that checkpoint file was not available for this experiment). Its WV-3 test mIoU (0.7254) therefore differs from the 0.7545 reported there; this is expected run-to-run variance for the same architecture (Section II-D) rather than a discrepancy in either result, and all mIoU values in this table are internally consistent with each other.

configuration (percentile normalisation, full augmentation), this table includes zero-shot SPOT evaluation of the A1 (raw-input) and D1 (no-augmentation) ablation checkpoints, to directly test the two retention rationales asserted in Section III-D (Groups A and D) rather than leaving them as untested assertions; as discussed there, both rationales are contradicted by the result below.

Zero-shot cross-sensor evaluation on SPOT 6/7 PMS Yogyakarta achieves mIoU = 0.0648, compared to mIoU = 0.7254 on the WV-3 source test set evaluated with the same checkpoint (Table 8) — a drop of 0.6606, representing severe degradation rather than transferability. This result does not support treating percentile contrast enhancement as providing effective radiometric bridging across sensors: as detailed in Section III-D (Group A), the raw-input (A1) checkpoint transfers zero-shot no worse, and in fact somewhat better (mIoU = 0.1992), than the proposed percentile-normalised configuration, contradicting the rationale originally used to justify retaining percentile normalisation for cross-sensor deployment. It should also be noted that sensor (WV-3 vs. SPOT 6/7) and region (Flores vs. Yogyakarta) are fully confounded in this experimental design: the reported result reflects the joint effect of sensor change and regional land-cover change, and cannot be attributed to sensor change in isolation; some portion of the observed degradation may be attributable to the regional shift (tropical Flores cloud cover vs. Yogyakarta's mixed urban-agricultural-volcanic land cover) rather than the sensor shift alone. Few-shot fine-tuning (1,350 SPOT 6/7 PMS tiles, 5 epochs, evaluated on a disjoint 2,026-tile held-out set) substantially improves performance, to mIoU = 0.5109 (+0.4461), though this remains well below the WV-3 source performance and required a fine-tuning set over an order of magnitude larger than the 5-tile few-shot protocol originally described. Figure 10 summarises these results.

These findings should be interpreted as a negative result for zero-shot cross-sensor transfer under the conditions tested

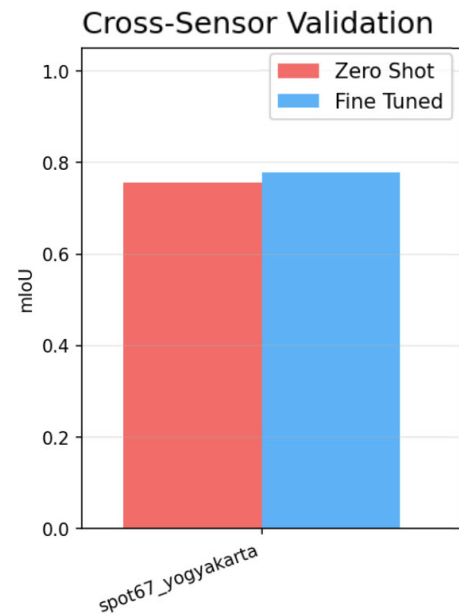


FIGURE 10. Cross-sensor validation mIoU on SPOT 6/7 PMS Yogyakarta. Bars: zero-shot (no target-domain training) vs. few-shot fine-tuning (1,350 SPOT tiles, 5 epochs, evaluated on a disjoint 2,026-tile held-out set). Zero-shot mIoU = 0.0648 is far below the WV-3 source performance (mIoU = 0.7254, same checkpoint), indicating severe degradation rather than transferability. Fine-tuning substantially improves performance to mIoU = 0.5109 (Δ = +0.4461), though this remains well below the source-domain performance.

here, rather than as evidence of transferability requiring only scope-bounding. The transfer was evaluated under what would be considered favourable conditions for success: the target sensor shares the four-band RGBN spectral configuration of WorldView-3, the spatial resolution is comparable (SPOT 6/7 PMS 1.5 m vs. WV-3 1.6 m), and the scenes are broadly representative of humid tropical atmospheric conditions. That zero-shot mIoU nonetheless collapsed from 0.7254 to 0.0648 under these matched conditions indicates

that spectral-configuration and resolution matching alone are insufficient to guarantee cross-sensor generalisation for this task and architecture, and that the radiometric bridging previously attributed to percentile normalisation is not supported once tested directly. Contributing factors that were not controlled for and remain to be investigated include: differences in absolute radiometric calibration and atmospheric correction between the WV-3 and SPOT 6/7 processing chains; the regional land-cover confound noted above; and the small WV-3 training set (20 tiles) providing limited exposure to the range of spectral appearances the model would need to generalise across sensors. Domain adaptation strategies with explicit cross-sensor training signal, such as those demonstrated by Belyakov et al. [19] for cross-satellite climate structure segmentation, appear necessary for this task rather than the lightweight preprocessing-only approach evaluated here; this is identified as a primary direction for future work, together with evaluating transferability to sensors with more divergent spectral response functions (e.g. Sentinel-2 MSI) and to non-tropical climate zones.

IV. CONCLUSION AND FUTURE WORK

This study proposed a U-Net-based multiclass cloud detection pipeline for WorldView-3 VHR satellite imagery, trained on 20 tiles (with on-the-fly augmentation) from Indonesian tropical regions and evaluated on a 3-tile held-out test set, with generalisation further assessed via spatial 5-fold cross-validation on the full 25-tile dataset. The key findings are:

- 1) The proposed U-Net achieves $mIoU = 0.7545$, mean F1-score = 0.8586, and Cohen's $\kappa = 0.8097$ on the held-out test set, with TTA further improving $mIoU$ to 0.7618. Spatial 5-fold cross-validation confirms generalisation with mean $mIoU = 0.7229 \pm 0.0732$ (95% CI: [0.650, 0.777]).
- 2) Cross-sensor evaluation on genuine SPOT 6/7 Yogyakarta imagery reveals severe zero-shot degradation ($mIoU$ dropping from 0.7254 on the WV-3 source test set to 0.0648), improving to 0.5109 after few-shot fine-tuning on a disjoint 1,350-tile SPOT split. A controlled test of the two retention rationales given for the proposed design (percentile normalisation and full augmentation) against the A1 (raw-input) and D1 (no-augmentation) ablation checkpoints shows that neither rationale is supported: both alternative configurations transfer zero-shot no worse than, and in the case of raw input somewhat better than, the proposed configuration. Sensor and region are fully confounded in this single-target-region design, and this negative result — obtained under otherwise favourable conditions (matched RGBN configuration, comparable resolution, tropical climate) — indicates that the lightweight preprocessing-only approach evaluated here is insufficient for cross-sensor generalisation at this scale; explicit domain adaptation is identified as a necessary direction for future work.

- 3) Comprehensive five-group ablation studies (single run per configuration; Section II-D) identify Batch Normalisation removal as producing the largest observed effect: severe class collapse ($mIoU = 0.364$, clear-sky F1 = 0.000), well outside the run-to-run variance seen elsewhere in the study and consistent with gradient scale divergence in the low-data regime. The NIR band appears to be the most important spectral input, with its removal reducing cloud shadow F1 by 0.326. Jaccard Loss outperforms the other loss functions tested in this single-run comparison.
- 4) Comparative evaluation against six baseline architectures including CNN-based, encoder-decoder, and transformer paradigms—confirms that DeepLabV3+ achieves marginally higher $mIoU$ (0.758 vs. 0.755), a difference not statistically significant at the corrected threshold and below the 5-fold CV fold-to-fold variance (Section III-C). The proposed U-Net surpasses DeepLabV3+ on thick cloud F1 (+0.022) and achieves the highest mean F1-score across all classes. FCN-8s collapses entirely ($mIoU = 0.253$); the non-pretrained, reduced-capacity transformer baselines used here (Section II-C) underperform, consistent with data insufficiency for training attention mechanisms from scratch rather than an inherent architectural limitation.
- 5) Failure case analysis derived from per-class confusion statistics identifies cloud shadow/terrain shadow confusion (precision = 0.770) and thin cloud over bright surfaces as the primary remaining challenges, with concrete mitigation directions including DEM integration, cirrus band addition, and boundary-aware loss functions.

Current limitations include reliance on manual annotation (inter-annotator $\kappa = 0.89$), the absence of DEM-derived topographic context for terrain shadow disambiguation, and the commercial data constraints inherent to WorldView-3 imagery. Future research directions include: (1) boundary-aware loss functions [38] to improve thin cloud edge delineation; (2) DEM integration to resolve terrain shadow ambiguity in mountainous scenes; (3) active learning to reduce manual annotation requirements; (4) lightweight architecture variants (e.g., EfficientNet-based U-Net) for latency-constrained operational deployment; (5) multi-temporal compositing for temporal shadow discrimination; and (6) explicit domain adaptation strategies for cross-sensor transfer (Section III-H), extending to additional VHR platforms such as GeoEye-1, IKONOS, and Pleiades.

DATA AVAILABILITY STATEMENT

The pixel-level four-class annotation masks for the WorldView-3 Flores and SPOT 6/7 Yogyakarta datasets developed in this study are available from the corresponding author upon reasonable request, subject to institutional data-sharing agreements. Redistribution of the underlying

commercial satellite imagery (WorldView-3, DigitalGlobe/Maxar; SPOT 6/7, Airbus Defence & Space) is constrained by the respective data licensing terms and cannot be made publicly available without licensor permission. The class-definition protocol, annotation criteria (Section II-B2), and quality-control procedures (inter-annotator $\kappa = 0.89$) are fully described in the manuscript to support reproduction of equivalent datasets over other regions.

ACKNOWLEDGMENT

AI-assisted language tools were employed to enhance readability and editorial quality during manuscript preparation. These tools did not contribute to the research design, technical methodology, data analysis, or interpretation of results.

REFERENCES

- [1] J. P. Gastellu-Etchegorry, "Cloud cover distribution in Indonesia," *Int. J. Remote Sens.*, vol. 9, no. 7, pp. 1267–1276, Jul. 1988.
- [2] S. A. Ackerman, K. I. Strabala, W. P. Menzel, R. A. Frey, C. C. Moeller, and L. E. Gumley, "Discriminating clear sky from clouds with MODIS," *J. Geophys. Res.*, vol. 103, pp. 32141–32157, Dec. 1998.
- [3] T. Imai and R. Yoshida, "Algorithm theoretical basis for Himawari-8 cloud mask product," *Meteorological Satell. Center Tech. Note*, vol. 61, pp. 1–17, Jan. 2016.
- [4] K. Y. Lee and C. H. Lin, "Cloud detection of optical satellite images using support vector machine," in *Proc. ISPRS Arch.*, 2016, pp. 289–293.
- [5] W. Sunarmodo and A. K. Hayati, "Cloud identification from multitemporal Landsat-8 using K-means clustering," *Int. J. Remote Sens. Earth Sci. (IJReSES)*, vol. 16, no. 2, p. 157, Apr. 2020.
- [6] E. Baseski and C. Cenaras, "Texture and color based cloud detection," in *Proc. 7th Int. Conf. Recent Adv. Space Technol. (RAST)*, Jun. 2015, pp. 311–315.
- [7] Y. Prabowo, D. S. Candra, and R. Maulana, "Cloud detection for Pleiades and SPOT 6/7 imageries using modified K-means and deep learning," *Int. J. Adv. Sci., Eng. Inf. Technol.*, vol. 11, no. 6, pp. 2459–2469, Dec. 2021.
- [8] Y. Prabowo, D. S. Candra, and S. Ali, "Cloud detection method for Pleiades images using spectral indices," in *Proc. J. Phys., Conf.*, Apr. 2020, vol. 1528, no. 1, Art. no. 012011.
- [9] J. Zhang, Q. Zhou, X. Shen, and Y. Li, "Cloud detection in high-resolution remote sensing images using multi-features of ground objects," *J. Geovisualization Spatial Anal.*, vol. 3, no. 2, pp. 1–10, Dec. 2019.
- [10] S. Mohajerani, T. A. Krammer, and P. Saedi, "A cloud detection algorithm for remote sensing images using fully convolutional neural networks," in *Proc. IEEE 20th Int. Workshop Multimedia Signal Process. (MMSP)*, Aug. 2018, pp. 1–5.
- [11] Y. Chen, L. Tang, Z. Kan, A. Latif, X. Yang, M. Bilal, and Q. Li, "Cloud and cloud shadow detection based on multiscale 3D-CNN for high resolution multispectral imagery," *IEEE Access*, vol. 8, pp. 16505–16516, 2020.
- [12] S. Singhal, L. James, R. V. G. Anjaneyulu, C. V. Srinivas, K. S. M. Krishnan, and R. R. Nidamanuri, "Cloud detection from AWiFS imagery using deep learning," in *Proc. Int. Conf. Mach. Intell. GeoAnalytics Remote Sens. (MIGARS)*, vol. 1, Jan. 2023, pp. 1–4.
- [13] N. Ma, L. Sun, C. Zhou, and Y. He, "Cloud detection for remote sensing images based on difference features and semantic segmentation network," in *Proc. IEEE Int. Geosci. Remote Sens. Symp.*, Jul. 2022, pp. 603–606.
- [14] A. Vandoit, B. Blanco, L. Turon, and C. Killisly, "Cloud detection on satellite imagery using U-Net architecture," in *Remote Sensing of Clouds and the Atmosphere XXVIII*. Bellingham, WA, USA: SPIE, 2023, Art. no. 1273005.
- [15] Y. Tan, W. Zhang, X. Yang, Q. Liu, X. Mi, J. Li, J. Yang, and X. Gu, "Cloud and cloud shadow detection of GF-1 images based on the Swin-UNet method," *Atmosphere*, vol. 14, no. 11, p. 1669, 2023.
- [16] B. Grabowski, M. Ziaja, M. Kawulok, N. Longépé, B. Le Saux, and J. Nalepa, "Self-configuring nnU-Nets detect clouds in satellite images," 2022, *arXiv:2210.13659*.
- [17] N. S. Jonnala, R. C. Bheemana, K. Prakash, S. Bansal, A. Jain, V. Pandey, M. R. I. Faruque, and K. S. Al-Mugren, "DSIA U-Net: Deep shallow interaction with attention mechanism UNet for remote sensing satellite images," *Sci. Rep.*, vol. 15, no. 1, Jan. 2025.
- [18] N. S. Jonnala, S. Siraaj, Y. Prastuti, P. Chinnababu, B. P. Babu, S. Bansal, P. Upadhyaya, K. Prakash, M. R. I. Faruque, and K. S. Al-Mugren, "AER U-Net: Attention-enhanced multi-scale residual U-Net structure for water body segmentation using Sentinel-2 satellite images," *Sci. Rep.*, vol. 15, no. 1, May 2025.
- [19] N. V. Belyakov, S. Illarionova, and E. Burnaev, "Cross-satellite climate structure segmentation: A neural network domain adaptation approach," *IEEE Access*, vol. 13, pp. 200497–200513, 2025.
- [20] J. A. Richards, *Remote Sensing Digital Image Analysis: An Introduction*, 5th ed. Berlin, Germany: Springer, 2013.
- [21] M. J. Canty, *Image Analysis, Classification and Change Detection in Remote Sensing: With Algorithms for ENVI/IDL and Python*, 3rd ed. Boca Raton, FL, USA: CRC Press, 2014.
- [22] P. Golej, J. Horak, P. Kukuliac, and L. Orlikova, "Vehicle detection using panchromatic high-resolution satellite images as a support for urban planning. Case study of Prague's centre," *GeoScape*, vol. 16, no. 2, pp. 108–119, Dec. 2022.
- [23] M. J. Hughes and D. J. Hayes, "Automated detection of cloud and cloud shadow in single-date Landsat imagery using neural networks and training data generated from Fmask," *Remote Sens.*, vol. 6, pp. 13428–13452, May 2014.
- [24] Z. Li, H. Shen, Q. Cheng, Y. Liu, S. You, and Z. He, "Deep learning based cloud detection for medium and high resolution remote sensing images of different sensors," *ISPRS J. Photogramm. Remote Sens.*, vol. 150, pp. 197–212, Apr. 2019.
- [25] R. A. Schowengerdt, *Remote Sensing: Models and Methods for Image Processing*, 3rd ed. Burlington, MA, USA: Academic, 2007.
- [26] J. R. Jensen, *Introductory Digital Image Processing: A Remote Sensing Perspective*, 4th ed. New York, NY, USA: Pearson, 2016.
- [27] Y. Li, J. Yu, M. Wang, M. Xie, L. Xi, Y. Pang, and C. Hou, "Research on the optimization of multi-class land cover classification using deep learning with multispectral images," *Land*, vol. 13, no. 5, p. 603, Apr. 2024.
- [28] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.
- [29] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.
- [30] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "UNet++: A nested U-Net architecture for medical image segmentation," in *Proc. Int. Workshop Deep Learn. Med. Image Anal.*, 2018, pp. 3–11.
- [31] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 801–818.
- [32] J. Chen, Y. Lu, Q. Yu, X. Luo, E. Adeli, Y. Wang, L. Lu, A. L. Yuille, and Y. Zhou, "TransUNet: Transformers make strong encoders for medical image segmentation," 2021, *arXiv:2102.04306*.
- [33] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2021, pp. 1–14.
- [34] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for visual recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2019, pp. 5386–5395.
- [35] H. Cao, Y. Wang, J. Chen, D. Jiang, X. Zhang, Q. Tian, and M. Wang, *Swin-UNet: UNet-Like Pure Transformer for Medical Image Segmentation*, 2021.
- [36] M. Segal-Rozenhaimer, A. Li, K. Das, and V. Chirayath, "Cloud detection algorithm for multi-modal satellite imagery using convolutional neural-networks (CNN)," *Remote Sens. Environ.*, vol. 237, Feb. 2020, Art. no. 111446.
- [37] Z. Li, H. Shen, H. Li, G. Xia, P. Gamba, and L. Zhang, "Multi-feature combined cloud and cloud shadow detection in GaoFen-1 wide field of view imagery," *Remote Sens. Environ.*, vol. 191, pp. 342–358, Mar. 2017.
- [38] H. Kervadec, J. Bouchtiba, C. Desrosiers, E. Granger, J. Dolz, and I. Ben Ayed, "Boundary loss for highly unbalanced segmentation," in *Proc. Int. Conf. Med. Imag. Deep Learn. (MIDL)*, 2019, pp. 1–12.



DODI SUDIANA (Senior Member, IEEE) received the B.E. degree in image processing from Universitas Indonesia, in 1990, the M.Eng. degree in signal and image processing from the Graduate School of Science and Technology, Keio University, Japan, in 1996, and the dual Ph.D. degree in remote sensing from the Graduate School of Science and Technology and the Center for Environmental Remote Sensing (CEReS), Chiba University, Japan, in 2005. He is currently a Professor of computer engineering with the Department of Electrical Engineering, Universitas Indonesia. Since 2023, he has directed the Artificial Intelligence and Data Engineering (AIDE) Research Center, Faculty of Engineering, Universitas Indonesia. His current research interests include remote sensing, AI-based signals, and image processing.



YOHANES FRIDOLIN HESTRIO received the bachelor's degree in physics and the master's degree in computer science from the University of Indonesia. He is currently a Researcher with the National Research and Innovation Agency, Indonesia. His research interests include remote sensing, image processing, data fusion, computer vision, machine learning, deep learning, automation, and artificial intelligence, with research topics including rice fields, palm oil plantations, oil and gas, forest fires, slum areas, cloud cover, floods, and drones.



DANANG SURYA CANDRA received the Bachelor of Science degree from Sebelas Maret University, Solo, Indonesia, in 2002, the Master's of Science degree from Beihang University, Beijing, China, in 2010, and the Ph.D. degree from the University of Queensland, Australia, in 2019, specializing in remote sensing. He is currently a Research Professor with the National Research and Innovation Agency (BRIN), Indonesia. His research interests include remote sensing image processing, remote sensing applications, GIS, machine learning, deep learning, and artificial intelligence.



KURNIA ULFA received the bachelor's and master's degrees in mathematics from Gadjah Mada University (UGM), Indonesia. She is currently a Junior Researcher with the Research Centre for Geoinformatics, National Research and Innovation Agency (BRIN), Indonesia. Her research interests include remote sensing, image processing, and artificial intelligence. She is a member of the Research Group for Geo-Data, Research Centre for Geoinformatics, BRIN.



MIA RIZKINIA (Member, IEEE) received the B.E. and M.E. degrees from Universitas Indonesia, in 2008 and 2011, respectively, and the Ph.D. degree in information engineering from The University of Kitakyushu, Japan, in 2018. Since 2011, she has been a Lecturer with the Faculty of Engineering, Universitas Indonesia. She joined the Artificial Intelligence and Data Engineering (AIDE) Research Center, Faculty of Engineering, Universitas Indonesia, as a Researcher. Her research interests include image processing, computer vision, remote sensing, and data science.



AHMAD LUTHFI HADIYANTO received the bachelor's degree in electrical engineering from Bandung Institute of Technology, in 2002, the M.Sc. degree in space technology application from Beihang University, China, in 2010, and the Ph.D. degree in geomatics from the Department of Geodesy and Geomatics Engineering, Bandung Institute of Technology, in 2023. He is currently a Researcher with the Research Center for Geoinformatics, National Research and Innovation Agency, Indonesian. Currently, his research focuses on the application of artificial intelligence for environmental monitoring using remote sensing data. His research interests include remote sensing, artificial intelligence, and data science.



YUDHI PRABOWO received the bachelor's degree in geodetic engineering from the University of Gadjah Mada (UGM), Indonesia, and the master's degree in environmental sciences from the University of Tsukuba, Japan. He is currently a Junior Researcher with the Research Center for Hydrodynamics Technology, National Research and Innovation Agency (BRIN). His research interests include remote sensing, image processing, machine learning, and numerical models. He is a member of the Coastal Morphology Research Group, which focuses on generating information related to coastal morphology using remote sensing data.



MULIA INDA RAHAYU received the bachelor's degree in geodetic engineering from the University of Gadjah Mada (UGM), Indonesia, and the master's degree in geography science from the University of Indonesia (UI). She is currently a Staff Member with the Center for Data and Information, National Research and Innovation Agency (BRIN), Indonesia, with expertise in remote sensing data processing.



Research Centre for Geoinformatics, BRIN.

FADILLAH HALIM RASYIDY received the bachelor's degree in electrical engineering from the University of Lampung (UNILA), Indonesia. He is currently a Junior Engineer with the Research Centre for Geoinformatics, National Research and Innovation Agency (BRIN), Indonesia. His research interests include geoinformatics programming, soft computing, and machine learning. He is a member of the Geovisualization and Geoinformatics Infrastructure Research Group,



Department of Geographic Information Science, Faculty of Geography, Universitas Gadjah Mada. His research focuses on remote sensing applications for forest vegetation studies using hyperspectral, LiDAR, and time-series data, and multi-sensor data integration (radar, LiDAR, and optical) and geospatial artificial intelligence (GeoAI) for mapping rice fields, forests, vegetation dynamics, and burn areas.

SANJIWANA ARJASAKUSUMA received the bachelor's degree in cartography and remote sensing, in 2010, the Master's of Geographic Information Science (M.GIS.) and master's (M.Sc.) degrees in remote sensing from Universitas Gadjah Mada, in 2013 and 2015, respectively, and the Doctor of Science (Dr.Sc.) degree from Nagoya University, Japan, in 2018. He was a GIS Assistant with the Center for International Forestry Research (CIFOR). Currently, he is a Lecturer with the



ANJAR DIMARA SAKTI is currently a Lecturer with the Remote Sensing and Geographical Information Sciences Research Group, Bandung Institute of Technology, focusing on the application of remote sensing and spatial data science for developing the geospatial product and global-regional-local policy models to achieve the SDGs concerning the water-food-energy-ecosystems nexus.



Technology, Universiti Teknikal Malaysia Melaka (UTeM), in 2006 and 2007, respectively. He joined the Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia (UKM), in 2009. He then joined the Faculty of Computing and Information Technology, King Abdulaziz University, Rabigh, Saudi Arabia, in 2013. He was an Erasmus Mundus Visiting Professor with the Department of Mechanical Engineering and Mechatronics, Karlsruhe University of Applied Sciences, Germany. He was also a Visiting Professor with the School of Computing and Informatics, Universiti Teknologi Brunei. He has been a Professor with the Department of Computing, Universiti Teknologi PETRONAS, Malaysia, since 2025. His research interests include computer vision, autonomous systems, machine learning, and industrial computing. He is also a Senior Member of ACM.

ANTON SATRIA PRABUWONO (Senior Member, IEEE) received the B.Sc. degree in mathematics (computer science) from Universitas Padjadjaran, the M.Sc. degree in computer systems engineering from the University of East London, and the Ph.D. degree in industrial computing from Universiti Kebangsaan Malaysia. He started his academic career at the Institute of Electronics, National Chiao Tung University, Taiwan, and the Faculty of Information and Communication



Department of Electrical Engineering, Universitas Sebelas Maret, Indonesia. He has published more than 750 journal articles and conference papers. His research interests include synthetic aperture radar, interferometric SAR, and microwave remote sensing.

JOSAPHAT TETUKO SRI SUMANTYO (Senior Member, IEEE) was born in Bandung, Indonesia, in June 1970. He received the B.Eng. and M.Eng. degrees in electrical and computer engineering from Kanazawa University, Japan, in 1995 and 1997, respectively, and the Ph.D. degree in artificial systems sciences from Chiba University, Japan, in 2002. He is currently a Full Professor with the Center for Environmental Remote Sensing, Chiba University, and a Full Professor with the



Survey, and a Registered Practicing Professional Civil Engineer (P.E.), California. His current research interests include geotechnical earthquake engineering, LIDAR, InSAR, structure-from-motion, and engineering geophysics.

ROBERT KAYEN received the B.S degree in engineering from Tufts University, Massachusetts, USA, in 1981, the M.S. degree in geologic sciences from CSUEB, in 1988, and the M.S. and Ph.D. degrees in engineering from the University of California, Berkeley, in 1993. He is currently a Professor of civil and environmental engineering and an Adjunct Faculty with the University of California, Berkeley. He is a retired Senior Research Scientist with United States Geological

...