

Cloud Segmentation for Optical Remote Sensing Satellite Imagery Using a Lightweight U-Net Architecture

Mohsn Alawi Alsaidi and Hatem M. Keshk*

Cite <https://doi.org/10.64589/juri/231545>

Submitted: November 26, 2025 Revised: May 3, 2026 Accepted: May 14, 2026

ABSTRACT

Cloud segmentation is a prerequisite for reliable analysis and cloud removal workflows in optical remote sensing imagery. In this study, we investigated a lightweight U-Net for binary cloud mask generation on the 38-Cloud dataset utilizing an EfficientNet-B0 encoder adapted for dense prediction. The network was trained using the AdamW optimizer and binary cross-entropy under a short baseline training schedule, resulting in an F1-score of 0.9517, precision of 0.9467, recall of 0.9567, IoU of 0.9079, and accuracy of 0.9698, while using 10.6 million parameters. Compared with heavier segmentation models, the proposed architecture offers a favorable accuracy-to-complexity tradeoff and remains practical for resource-constrained remote sensing workflows. Our findings demonstrate that a carefully optimized encoder-decoder design can deliver competitive cloud segmentation performance without relying on computationally expensive attention or transformer modules.

Keywords: cloud segmentation, satellite imagery, semantic segmentation, EfficientNet-B0, U-Net, deep learning, remote sensing, lightweight model

1. INTRODUCTION

1.1. Background and Motivation. Remote sensing utilizes various sensor platforms to capture data remotely. Such platforms include traditional aircraft, unmanned aerial vehicles (UAVs), satellites, and CubeSats, which are deployed with different types of sensors. The most common sensor types include digital cameras, multispectral cameras, and infrared sensors, which capture images using electromagnetic radiation. Satellite imagery is an important branch of remote sensing, which is ideal for capturing large areas of the Earth, whereas aircraft and drones are typically operated to cover limited regions. Remote sensing applications include land cover monitoring, disaster response and recovery, agriculture, geology and natural resource management, military surveillance, and oceanography, emphasizing the importance of this field. However, a major challenge associated with satellite imagery is the frequent presence of clouds. Clouds block surface information, reduce image quality, and limit the usefulness of remote sensing applications. Accordingly, the development of accurate cloud segmentation approaches is crucial for improving the quality and accuracy of satellite data^{1,2}.

1.2. Research Objectives.

1.2.1. Problem statement. Satellite imagery is often affected by geometric and radiometric distortions, as well as atmospheric disruptions, including clouds, aerosol particles, dust particles, contaminants, and gases. Sensor response variability, calibration-related errors, nonuniform illumination, and image sensor noise may also reduce imaging quality. At any given time,

approximately 67% of the Earth is covered by clouds¹. Accordingly, many methods have been developed for cloud removal in satellite imagery, ranging from traditional thresholding methods to machine learning and deep learning approaches, which are considered automated methodologies. In particular, deep learning automates the training process and does not require manual threshold selection or handcrafted features³. Although pixel-level manual annotation for cloud mask production is considered the most well-established and effective method, the large amount of data required for labeling makes this approach impractical for real-world applications^{3,4}.

1.2.2. Objectives. The main objectives of this study are as follows:

- To design a lightweight U-Net with an EfficientNet-B0 encoder that reduces computational complexity while maintaining strong segmentation performance.
- To compare the proposed model with reported 38-Cloud baselines and evaluate the tradeoff between predictive performance and parameter efficiency.

1.2.3. Contributions.

1. **Model Design:**
A lightweight U-Net segmentation model was developed utilizing EfficientNet-B0 as an encoder to balance representation quality and computational efficiency.
2. **Encoder Simplification:**
The EfficientNet-B0 encoder was simplified to focus on segmentation, rather than image classification, enabling us to

examine the extent to which a classical U-Net pipeline can be leveraged before adding heavier attention-based or nested components.

3. Performance Validation:

The proposed model was trained and evaluated on the 38-Cloud dataset and achieved competitive results across standard metrics, including accuracy, precision, recall, F1-score, IoU, and Cohen's kappa.

4. Reproducible Framework:

Our workflow followed a reproducible notebook-based pipeline utilizing publicly available 38-Cloud preprocessing resources, making our implementation accessible for future benchmarking and extension.

2. LITERATURE REVIEW

2.1. Traditional Cloud Detection Approaches.

2.1.1. Threshold-based methods. Threshold-based cloud detection methods were among the earliest operational approaches for optical satellite imagery. Such methods typically rely on spectral tests, histogram-based thresholds, thermal cues, or cirrus band information to distinguish cloudy and clear pixels. Foundational operational systems, such as Fmask and related rule-based masking pipelines, demonstrate that carefully designed thresholds can provide strong baselines, especially when sensor-specific knowledge is available. However, the performance of such approaches may degrade under conditions with bright surfaces, thin clouds, haze, or complex backgrounds.

2.1.2. Machine learning. Classical machine learning methods for cloud segmentation include support vector machines, random forests, classification and regression trees, k-nearest neighbors, naïve Bayes, and shallow neural networks. These methods learn decision boundaries from labeled samples and often improve upon fixed thresholding; however, they still depend strongly on handcrafted spectral or texture descriptors. As a result, these methods may struggle to generalize when cloud appearance changes across sensors, land cover types, seasons, and illumination conditions.

2.2. Deep Learning for Cloud Segmentation. Deep learning approaches learn hierarchical spatial features directly from imagery and have become the dominant paradigm for cloud masking. Convolutional encoder-decoder models, particularly U-Net and its variants, are well-suited to pixel-wise prediction because they combine contextual semantics from deep layers with fine spatial details from skip connections. Several studies have further improved performance by incorporating attention modules, nested skip pathways, transformer encoders, and hybrid spectral-spatial designs, although these gains often result in higher computational costs.

2.3. EfficientNet. EfficientNet is a lightweight backbone family designed for image classification and based on the compound scaling of depth, width, and resolution^{5,9}. EfficientNet-B0 is the smallest member of the original series and utilizes mobile inverted bottleneck convolution (MBConv) blocks with depth-wise separable convolutions and squeeze-and-excitation modules⁶. Because this model offers a favorable balance between

Table 1. Landsat 8 data: bands, wavelengths, and spatial resolution

Band	Name	Spatial resolution (m)	Wavelength (μm)
1	Ultra Blue	30 m	0.435–0.451
2	Blue	30 m	0.452–0.512
3	Green	30 m	0.533–0.590
4	Red	30 m	0.636–0.673
5	Near Infrared (NIR)	30 m	0.851–0.879
6	SWIR 1	30 m	1.566–1.651
7	SWIR 2	30 m	2.107–2.294
8	Panchromatic	15 m	0.503–0.676
9	Cirrus	30 m	1.363–1.384
10	Thermal Infrared (TIRS) 1	100 m	10.60–11.19
11	Thermal Infrared (TIRS) 2	100 m	11.50–12.51

feature quality and parameter count, it is a reasonable starting point for lightweight segmentation research, given that its classification-oriented tail is carefully adapted to dense prediction.

2.4. Dataset and Preprocessing.

2.4.1. Dataset. The dataset used in this study was obtained from Landsat 8, which launched in 2013. Landsat 8 contains nine bands captured by an operational land imager instrument and two thermal spectral bands captured by a thermal infrared sensor (TIRS) instrument, covering various wavelengths. The cloud cover assessment system and FMask algorithm were developed to provide a quality assessment (QA) band for Landsat 8 data^{7,21}. Default ground-truth results for clouds and snow were obtained from these data⁷. Table 1 lists the bands, wavelengths, and spatial resolutions in Landsat 8¹⁸.

2.4.2. Preprocessing. A snow/ice elimination threshold-based method was adopted to refine the cloud ground-truth labels derived from the original Landsat 8 data. Specifically, the QA band provided by Landsat 8 was used to divide images into snow, cloud, and clear regions. The average gradient magnitudes in these three regions were determined by calculating the gradient magnitude for each pixel. The snow-region average exhibited considerable differences from the remaining images. In particular, the second band exhibited the highest relative difference in average gradient magnitudes and was utilized for the snow/ice removal framework. Binary snow masks were produced by applying a global threshold to separate the high-intensity pixels. Utilizing this threshold-based method, we upgraded the 38-Cloud dataset with corrected binary cloud masks with higher accuracy by removing the snow-labeled regions present in both the default cloud ground-truth labels and Landsat 8 QA band data⁷. Subsequently, five incorrect ground-truth labels were replaced to enhance the accuracy of the dataset⁸.

In the Kaggle 38-Cloud dataset code, a spectral set of red, green, blue, and near-infrared channels was combined to produce

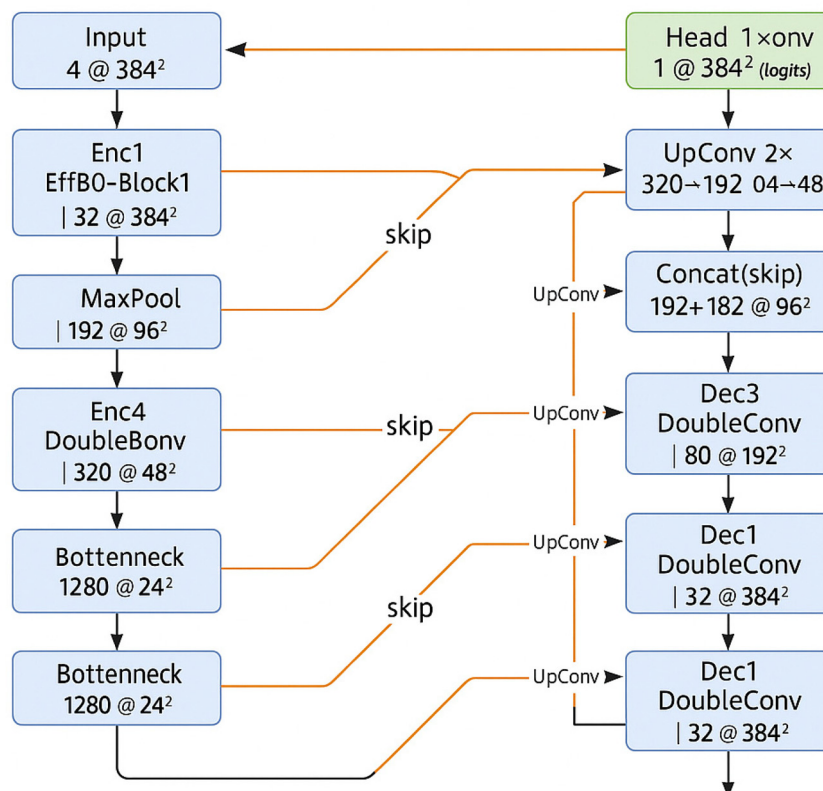


Figure 1. Architecture of the proposed U-Net with an EfficientNet-B0 encoder and lightweight decoder path

Table 2. Summary of evaluation metrics for the proposed method

Accuracy	Precision	Recall	F1 Score	IoU	Kappa	Parameters
0.9698	0.9467	0.9567	0.9517	0.9079	0.9298	10.6 M

four spectral input images, which were resized to 384×384 pixels and normalized into pixel intensities. The resulting four spectral image filters enhance deep learning by providing additive informative features.

2.5. Summary and Gap Identification. The literature follows two parallel trends in cloud segmentation research: operational threshold-based systems, such as Fmask and the Maccs-Atcor joint algorithm, remain important because they are efficient and interpretable, whereas recent deep learning models achieve higher accuracy by increasing architectural complexity. These trends leave a practical gap for methods that remain lightweight, reproducible, and competitive on benchmark datasets. The proposed EfficientNet-B0 U-Net is intended to address this gap by retaining a classical encoder-decoder structure while reducing unnecessary model overhead.

3. METHODOLOGY

3.1. Encoder-decoder U-Net. The proposed model follows the standard U-Net encoder-decoder paradigm. In the encoder, feature maps are progressively downsampled to enlarge the receptive field and capture higher-level semantic context. In the decoder, upsampling layers reconstruct the original spatial

resolution, while skip connections transfer fine-grained localization cues from the encoder to the corresponding decoder stages. This combination is well-suited to cloud segmentation because

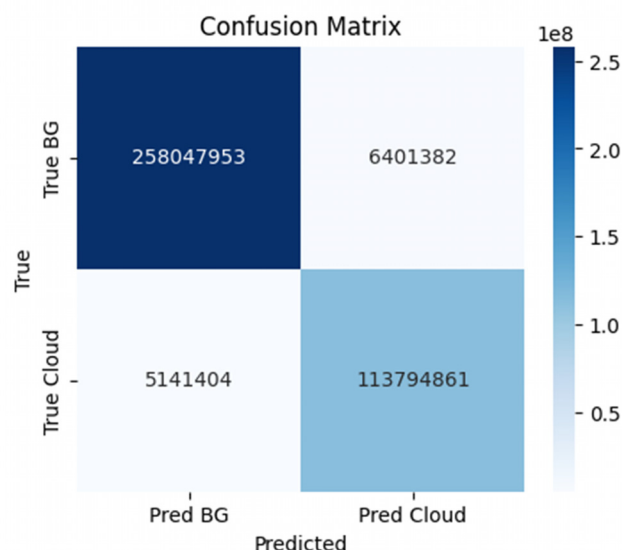


Figure 2. Confusion matrix for the proposed method, calculated based on 2600 samples gathered from the validation split data

cloud boundaries can be diffused at the scene level, but still require accurate pixel-level delineation.

EfficientNet-B0 was selected as the encoder architecture because it provides strong feature extraction capabilities at a relatively low computational cost. However, the original backbone was designed for image classification, rather than dense prediction. Therefore, the classification head was removed, and the feature hierarchy was coupled with a U-Net-style decoder. The architectural goal was not to discard the core EfficientNet-B0 representation, but to adapt it into a segmentation pipeline with compatible skip connections, stable spatial resolutions, and a lightweight decoding path. Figure 1 illustrates the workflow and channel transitions used by the final network.

3.2. Training Configuration.

3.2.1. Training strategy. The 38-Cloud dataset contains 8,400 training patches and 9,201 testing patches constructed from RGB and near-infrared imagery. In this study, the available training set was randomly divided into 5,400 samples for model fitting and 2,600 samples for validation, and the inputs were converted into tensors following resizing and normalization.

The model was trained using shuffled mini-batches for 15 epochs, consistent with the learning curves reported later in this paper.

A relatively short training schedule was adopted to establish a lightweight baseline with limited notebook computing resources and to examine whether the proposed model can converge quickly without heavy tuning. The loss and validation curves indicated continual improvement with partial stabilization by the final epochs, suggesting that the selected schedule was sufficient for a strong baseline, but not necessarily for full convergence. It should be noted that the results reported here correspond to a single training-validation split. The variance across repeated runs was not measured, but should be included in future studies.

3.2.2. Loss function. Binary cross-entropy was used as the optimization objective because the target task involves binary semantic segmentation. This loss function penalizes disagreement between the predicted cloud probabilities and ground-truth masks at the pixel level and is simple to optimize utilizing probabilistic outputs.

$$E = \sum_{x \in \Omega} w(x) \log(p_{l(x)}(x)) \quad (1)$$

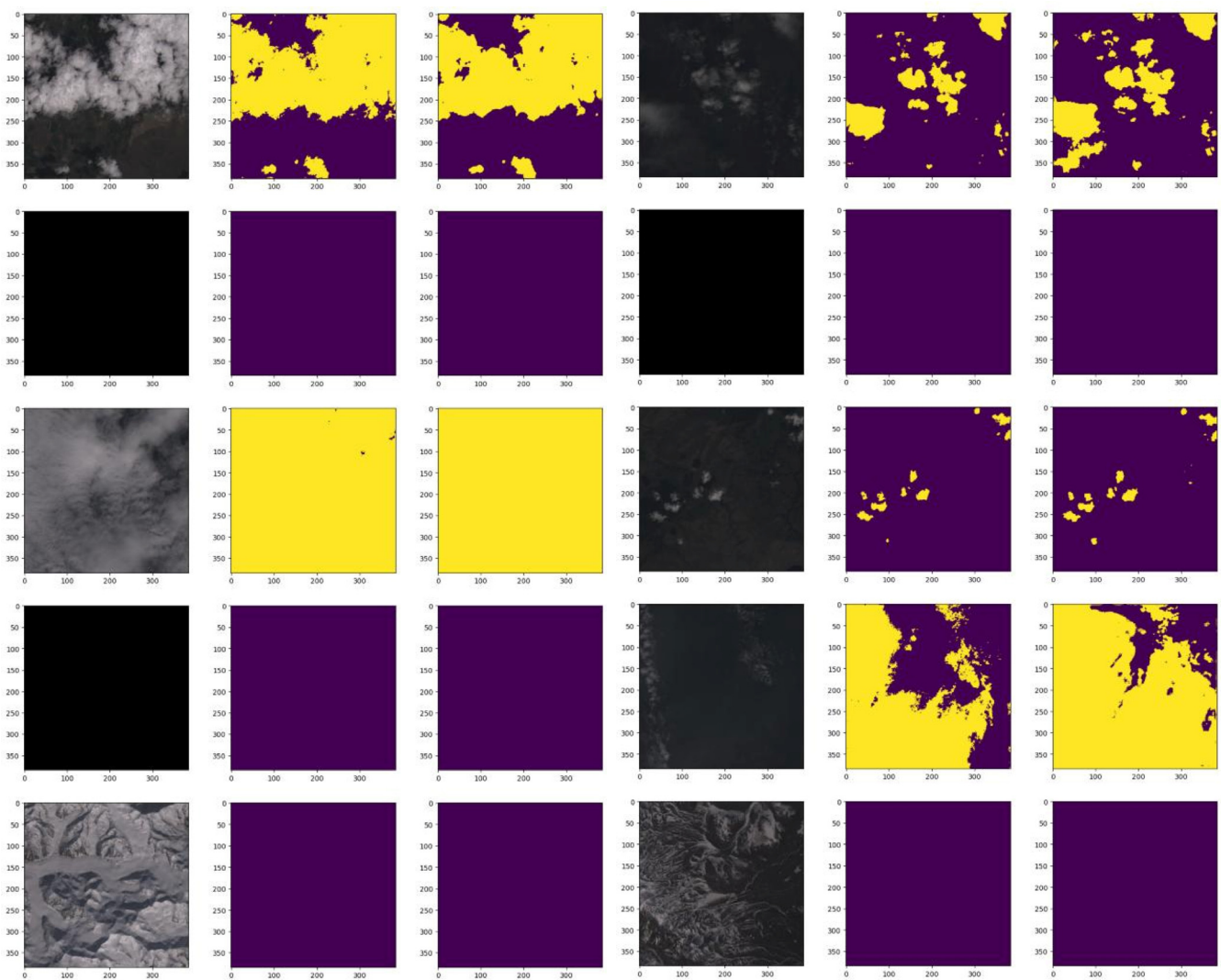


Figure 3. Representative image samples. For each image, the first column represents the input image in RGB, the second column represents the ground-truth mask, and the third column represents the semantic mask generated by the proposed method

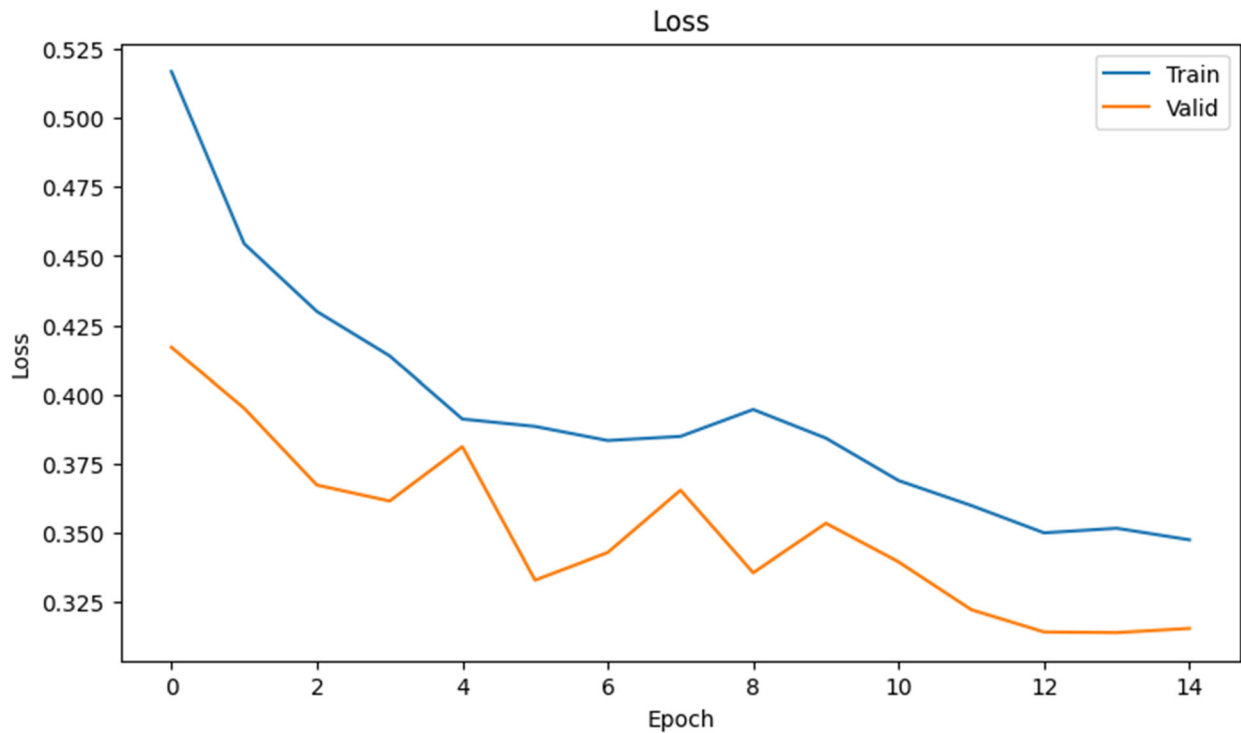


Figure 4. Training and validation loss across training epochs

3.2.3. Optimizer and learning rate. AdamW was used as the optimizer with a learning rate of 0.001. AdamW decouples weight decay from gradient updating and is commonly used to improve generalization compared with the standard Adam optimizer under the same training setup.

3.2.4. Hardware and software environment. All experiments were conducted on Google Colab using an NVIDIA T4 GPU and approximately 50 GB of RAM in an Ubuntu 22.04.4 LTS environment. Python 3.12, PyTorch, Pandas, scikit-learn, and matplotlib were used for implementation.

4. RESULTS AND DISCUSSION

4.1. Quantitative Evaluation.

4.1.1. Evaluation metrics. Our evaluation used the following standard segmentation metrics derived from the confusion matrix: accuracy, precision, recall, F1-score, IoU, and Cohen's kappa coefficient. Together, these measures quantify overall correctness, cloud detection sensitivity, false alarm control, class overlap quality, and agreement beyond chance. The metrics were calculated as shown in Eqs. (2) to (7). Table 2 summarizes the results of calculating each metric for the proposed method.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (2)$$

$$\text{Precision} = \frac{TP}{TP + FP} \quad (3)$$

$$\text{Recall (TPR)} = \frac{TP}{TP + FN} \quad (4)$$

$$\text{F1 score} = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \quad (5)$$

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

$$\kappa = \frac{(TP + TN) - \left[\frac{(TP + FP)(TP + FN) + (FN + TN)(FP + TN)}{(TP + TN + FP + FN)} \right]}{(TP + TN + FP + FN) - \left[\frac{(TP + FP)(TP + FN) + (FN + TN)(FP + TN)}{(TP + TN + FP + FN)} \right]} \quad (7)$$

4.1.2. Confusion matrix. The confusion matrix in Figure 2 provides a class-wise summary of model behavior. Most cloud pixels were classified correctly, and the largest residual errors were associated with the confusion between clouds and background pixels near challenging boundaries or visually ambiguous regions. This pattern is consistent with the qualitative examples presented later, with the model generally capturing the main cloud structures, but occasionally over-segmenting or under-segmenting thin or fragmented formations.

4.2. Qualitative Results. Figure 3 presents representative qualitative results. The first, second, and third columns contain the input satellite images, reference masks, and predicted masks, respectively. In most cases, the proposed model captured the dominant cloud extent and preserved the overall geometry of the cloud field. The errors primarily consist of small boundary mismatches, missed thin-cloud regions, or slight over-predictions in difficult background areas.

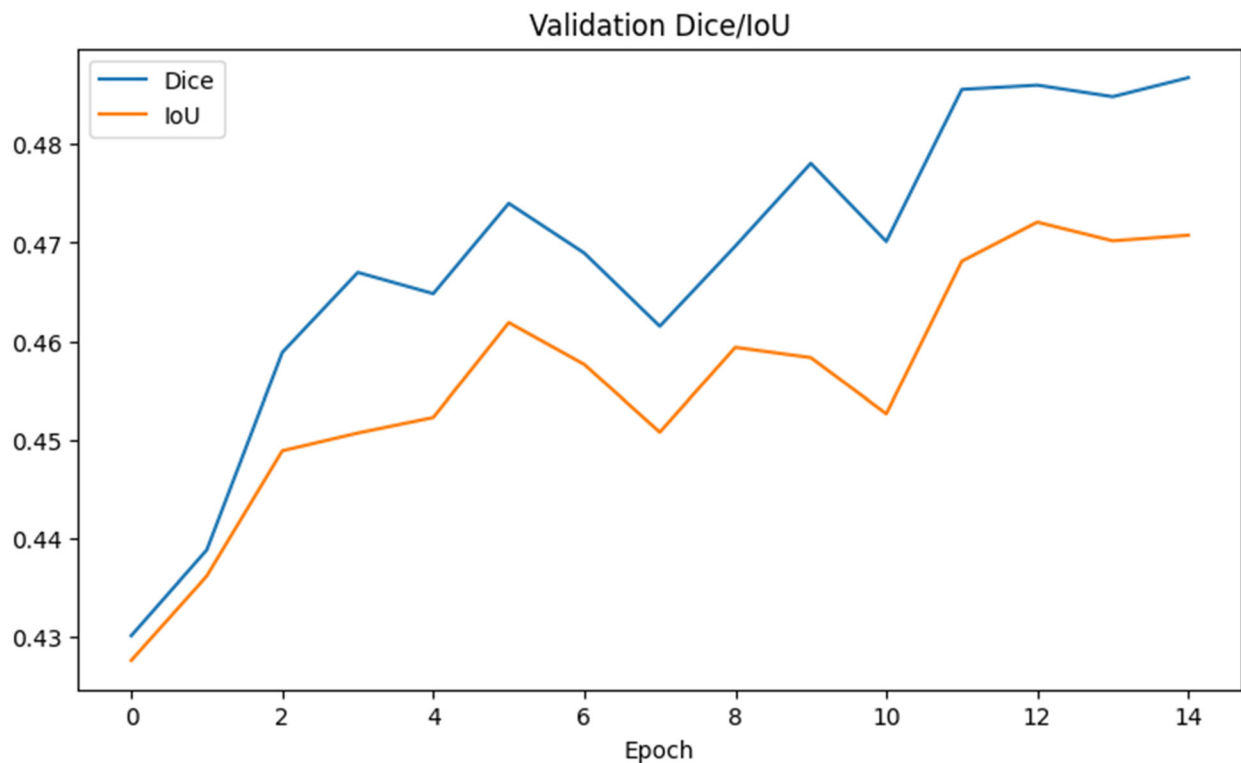


Figure 5. Validation Dice and IoU across training epochs

4.3. Model Training Performance.

4.3.1. Training and validation loss. As shown in Figure 4, the training and validation losses decreased overall with moderate fluctuations, and there was no evidence of severe divergence between the two curves. As shown in Figure 5, the validation Dice and IoU curves also improved across epochs, indicating that the network continued to learn useful segmentation features during training. As the curves continued to improve near the end of the schedule, longer training or repeated runs may continue to refine the final results.

4.4. Comparative Evaluation. As shown in Table 3, the proposed EfficientNet-B0 model achieved competitive performance relative to many reported 38-Cloud baselines while remaining substantially lighter than several attention-based, transformer-based, or heavily parameterized alternatives. Although the best F1 values were achieved by other methods, the proposed model offers a strong balance between accuracy, architectural simplicity, and parameter efficiency, which are important factors for reproducible experimentation and deployment in constrained settings.

Table 3. Comparisons between previous models and our proposed model

Method	Accuracy (%)	F1 (%)	Precision (%)	Recall (%)	Parameters
FCN ⁷	95.23	82.95	84.59	81.37	–
Fmask V3 ¹³	94.94	92.42	88.65	96.52	–
U-Net ⁷	95.05	91.91	96.15	88.02	7.07 M
*PSPNet ¹⁷	95.0	95.14	94.1	96.2	60.09 M
UNet+ + ¹⁶	95.9	96.25	96.7	95.8	7.79 M
*TransUNet ¹⁵	96.58	96.62	96.58	96.66	66.88 M
Cloud-AttU ¹⁴	97.05	93.96	97.16	91.30	430 M
Cloud-Net ⁸	95.86	93.23	97.6	89.23	32.9 M
Cloud-Net+ + ¹³	96.35	94.09	97.39	91.01	36.4 M
CUA-Net ¹²	96.72	89.82	88.58	91.10	–
RD-UNet ¹¹	96.81	91.91	90.80	93.04	359.0 M
SAT-UNet ¹⁰	96.65	96.69	96.17	97.23	9.78 M
EfficientNet-B0 U-Net (proposed model)	96.98	95.17	94.67	95.67	10.6 M

4.5. Limitations. Despite the excellent performance achieved by the proposed model on the 38-Cloud dataset, several limitations remain. First, the proposed method was evaluated using a single benchmark dataset, which limits its transferability to other sensors, terrains, and atmospheric conditions. Second, the training schedule was intentionally short, meaning the network may not have reached full convergence. Third, inputs were limited to the RGB and near-infrared bands, which may restrict the use of additional spectral cues. Fourth, reference masks were derived from algorithmic processing, rather than exhaustive expert pixel-level annotation, meaning a certain degree of label noise was likely present. Finally, repeated runs were not performed, so performance variance across random initializations and data splits remains unreported. To assess robustness further, future work will include multiple training runs with different random seeds and the calculation of mean $\pm$ standard deviation values. Due to computational constraints, this study represents a single-run baseline.

5. CONCLUSIONS

This paper presented a lightweight U-Net for cloud segmentation in optical remote sensing imagery utilizing an EfficientNet-B0 encoder. On the 38-Cloud dataset, the proposed model achieved F1-score, IoU, and accuracy values of 0.9517, 0.9079, and 0.9698, respectively, with only 10.6 million parameters. These results indicate that competitive cloud mask generation does not necessarily require heavy architectures. Instead, a carefully adapted classical encoder-decoder architecture can provide an effective balance between segmentation quality and computational efficiency. Future research should focus on repeated-run variance, extend evaluations to additional datasets and sensors, and examine whether longer training or richer spectral inputs further improve robustness.

AFFILIATIONS AND AUTHOR DETAILS

Undergraduate Author

Mohsn Alawi Alsaïdi – Department of Aerospace Engineering,
King Fahd University of Petroleum & Minerals, Dhahran
31261, Saudi Arabia; ☎ 0009-0009-9568-8752
Email: s202066080@kfupm.edu.sa

Corresponding Author

Hatem M. Keshk – Research Mentor, Postdoc, Interdisciplinary Research Center for Aviation and Space Exploration,
King Fahd University of Petroleum & Minerals, Dhahran
31261, Saudi Arabia; ☎ 0000-0002-8043-1227
Email: hatem.keshk@kfupm.edu.sa

ACKNOWLEDGEMENTS

The authors express their sincere gratitude to their advisors for their continuous guidance, encouragement, and support throughout this study. Appreciation is also extended to the Interdisciplinary Research Center for Aviation and Space Exploration

and the Department of Aerospace Engineering at King Fahd University of Petroleum and Minerals (KFUPM) for providing a supportive research environment.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

REFERENCES

- (1) Singh, R., Pal, M. & Biswas, M. Cloud Detection Methods for Optical Satellite Imagery: A Comprehensive Review. *Geomatics* **5**, 27 (2025).
- (2) Li, J., Cai, Y., Li, Q., Kou, M. & Zhang, T. A review of remote sensing image segmentation by deep learning methods. *International Journal of Digital Earth* **17**, (2024).
- (3) Wang, Z. *et al.* Deep Learning-Based Cloud Detection for Optical Remote Sensing Images: a survey. *Remote Sensing* **16**, 4583 (2024).
- (4) Rewhel, E. M. *et al.* Deep learning methods used in remote sensing images: a review. *Journal of Environmental & Earth Sciences* **5**, 33–64 (2023).
- (5) Elharrouss, O., Akbari, Y., Almadeded, N. & Al-Maadeded, S. Backbones-review: Feature extractor networks for deep learning and deep reinforcement learning approaches in computer vision. *Computer Science Review* **53**, 100645 (2024).
- (6) Amjoud, A. B. & Amrouch, M. Object Detection Using Deep Learning, CNNs and Vision Transformers: A review. *IEEE Access* **11**, 35479–35516 (2023).
- (7) Mohajerani, S., Krammer, T. A. & Saeedi, P. Cloud detection algorithm for remote sensing images using fully convolutional neural networks. In *2018 IEEE 20th International Workshop on Multimedia Signal Processing (MMSP)* 1–6 (IEEE, 2018).
- (8) Mohajerani, S. & Saeedi, P. Cloud-Net: An end-to-end cloud detection algorithm for Landsat 8 imagery. In *2019 IEEE International Geoscience and Remote Sensing Symposium (IGARSS)* 1029–1032 (IEEE, 2019).
- (9) Tan, M. & Le, Q. EfficientNet: Rethinking model scaling for convolutional neural networks. In *Proceedings of the 36th International Conference on Machine Learning*, vol. 97 of *Proceedings of Machine Learning Research*, 6105–6114 (2019).
- (10) Aimyshev, D., Tulegenov, B., Smadyarov, B. & Nurzhakyp, D. GhalamClouds: Remote sensing images for Kazakhstan and clouds segmentation using SAT-UNet. *Advances in Space Research* **77**, 1–20 (2026).
- (11) Li, A., Li, X. & Ma, X. Residual dual U-shape networks with improved skip connections for cloud detection. *IEEE Geosci. Remote Sens. Lett.* **21**, 1–5 (2024).
- (12) Li, A., Yang, J. & Li, X. Cloud detection from remote sensing images by cascaded U-shape attention networks. In *Lecture Notes in Computer Science*, 155–166 (2023).
- (13) Mohajerani, S. & Saeedi, P. Cloud and cloud shadow segmentation for remote sensing imagery via filtered Jaccard loss function and parametric augmentation. *IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens.* **14**, 4254–4266 (2021).
- (14) Guo, Y., Cao, X., Liu, B. *et al.* Cloud detection for satellite imagery using attention-based U-Net convolutional neural network. *Symmetry* **12**, 1056 (2020).
- (15) Chen, J., Xie, Y., Abuduweili, J. G. *et al.* TransUNet: Transformers make strong encoders for medical image segmentation. *arXiv* **2102.04306** (2021).
- (16) Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N. & Liang, J. UNet++: A nested U-Net architecture for medical image segmentation.

In *Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support*, vol. 11045 of *Lecture Notes in Computer Science*, 3–11 (Springer, 2018).

(17) Zhao, Z., Shi, J., Qi, X. et al. Pyramid scene parsing network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* 6230–6239 (2017).

(18) U.S. Geological Survey. *Landsat 8 (L8) Data Users Handbook*. (U.S. Geological Survey, 2019). Available at: <https://www.usgs.gov/landsat-missions/landsat-8-data-users-handbook> (accessed 18 Nov 2025).

(19) I. Loshchilov and F. Hutter, “Decoupled weight decay regularization,” arXiv preprint [arXiv:1711.05101](https://arxiv.org/abs/1711.05101), 2017.

(20) Zhu, Z. & Woodcock, C. E. Object-based cloud and cloud shadow detection in Landsat imagery. *Remote Sensing of Environment* 118, 83–94 (2012).

(21) Zhu, Z., Wang, S. & Woodcock, C. E. Improvement and expansion of the Fmask algorithm: Cloud, cloud shadow, and snow detection for Landsats 4–7, 8, and Sentinel 2 images. *Remote Sensing of Environment* 159, 269–277 (2015).