Received: 15 November 2025 | Revised: 27 May 2026 | Accepted: 19 June 2026 | Published online: 13 August 2026

RESEARCH ARTICLE

Journal of Computational and Cognitive Engineering
2026, Vol. 00(00) 1–20
DOI: 10.47852/bonviewJCCE62028252

BON VIEW PUBLISHING

A Dynamic Confidence-Aware Mask Fusion Framework for Aerial Image Segmentation Using U-Net and U-Net++

Ahmad Qasim AlHamad[1,*], Ahmad Mousa Odat[2], Abdullah Ahmad Al Sokkar[3], Mohammed Abdallah Otair[4], Suhier Nadi Odah[5] and Omar Hussain Tarawneh[5]

[1] *Department of Business Administration, University of Sharjah, United Arab Emirates*

[2] *Faculty of Information Technology, Jadara University, Jordan*

[3] *Faculty of Business, Applied Science Private University, Jordan*

[4] *Faculty of Information Technology, Middle East University, Jordan*

[5] *Faculty of Information Technology, Amman Arab University, Jordan*

Abstract: Semantic segmentation of aerial and drone imagery remains a demanding task due to class imbalance, scale diversity, and the visual similarity between land-cover categories. Although encoder–decoder architectures such as U-Net and U-Net++ have demonstrated strong performance, reliance on a single model often limits generalization across heterogeneous datasets. This study introduces a Dynamic Confidence-Aware Mask Fusion (DCAMF) framework that reformulates model fusion as a pixel-wise, uncertainty-guided decision process. Instead of applying static averaging or conventional voting schemes, the proposed approach integrates heterogeneous U-Net and U-Net++ models equipped with multiple backbone encoders and assigns adaptive weights based on entropy-derived confidence measures. The framework was evaluated on two multi-class aerial datasets with differing levels of semantic complexity. Experimental findings indicate consistent improvements in mean Intersection over Union and overall pixel accuracy compared with individual architectures and traditional ensemble strategies. The gains were particularly evident in boundary regions and visually ambiguous areas, where model disagreement is common. By incorporating uncertainty modeling directly into mask-level aggregation, DCAMF enhances robustness without introducing architectural modifications or excessive computational overhead. The results demonstrate that confidence-aware decision fusion offers a practical and scalable pathway for improving segmentation reliability in real-world aerial imaging applications.

Keywords: semantic segmentation, model fusion, convolutional neural networks (CNNs)

1. Introduction

Semantic segmentation is a fundamental task in computer vision that assigns class labels to individual pixels within an image, thereby enabling detailed scene interpretation. Its applications span diverse domains, including remote sensing, environmental monitoring, precision agriculture, disaster management, urban planning, and autonomous navigation [1–4]. In aerial and drone imagery, semantic segmentation plays a critical role in extracting meaningful information from complex and high-resolution data to support decision-making processes in smart cities and geospatial intelligence [5].

Traditional image segmentation approaches—such as thresholding, clustering, and region growing—struggle to handle the variability and complexity of aerial imagery [6]. These methods are often sensitive to noise, illumination changes, and intra-class variability, leading to poor generalization. The introduction of deep learning and, more specifically, convolutional neural networks (CNNs) has revolutionized this field by enabling end-to-end learning of robust hierarchical features directly from raw data.

Among CNN-based architectures, U-Net [7] and its enhanced variant U-Net++ [8] have become prominent due to their encoder–decoder structures with skip connections that preserve spatial details while capturing semantic context. U-Net has demonstrated strong performance in biomedical imaging and remote sensing, while U-Net++ introduces nested skip pathways

***Corresponding author:** Ahmad Qasim AlHamad, Department of Business Administration, University of Sharjah, United Arab Emirates. Email: aalhamad@sharjah.ac.ae

that improve feature fusion and enhance segmentation accuracy across multiple scales [9].

Despite their success, several challenges remain. Deep CNNs are data-hungry and computationally expensive, requiring large-scale annotated datasets and powerful hardware resources [10]. Moreover, their performance is often constrained by the representational capacity of the backbone networks used in the encoder. In practice, single-model architecture struggles to generalize effectively when applied to heterogeneous aerial datasets with multiple classes and complex background structures [11].

To overcome these challenges, ensemble and fusion strategies have gained traction in recent years. By combining the strengths of multiple models, fusion approaches can reduce misclassification, improve generalization, and enhance robustness against dataset variability [12]. Fusion can be performed at different levels—feature level, decision level, or mask level—each offering trade-offs between computational efficiency and accuracy.

Unlike existing segmentation studies that primarily focus on improving individual network architectures or feature-level fusion, this work emphasizes confidence-driven decision-level fusion across heterogeneous segmentation models. To the best of our knowledge, this is among the first studies to systematically combine U-Net and U-Net++ variants using a unified mask-based confidence-aware fusion strategy for multi-class aerial imagery segmentation.

2. Related Work

2.1. Review methodology

To ensure a comprehensive and unbiased review, we conducted a summarized systematic literature review (SLR) following PRISMA-style guidelines [13]. The search focused on recent advances in semantic segmentation of aerial and drone imagery. Major academic databases were queried, including Scopus, IEEE Xplore, SpringerLink, and Elsevier ScienceDirect. Search terms combined keywords such as "semantic segmentation," "U-Net," "aerial images," "drone images," "fusion," and "deep learning." A total of 54 papers were initially retrieved. After screening titles/abstracts and removing duplicates, 23 studies were fully reviewed, and 9 representative studies were included in this synthesis.

2.2. Study categorization

The selected literature was categorized into four major groups:

1) CNN-based baselines

Early deep architectures such as FCN, SegNet, and DeepLab provided the foundation for semantic segmentation in remote sensing. They demonstrated strong contextual modeling but often underperformed on small-object classes.

2) U-Net and its variants

U-Net [7] and U-Net++ [8] have been widely adopted. Recent studies augmented these models with attention mechanisms, dense connections [14], or lightweight backbones [15] to balance accuracy and efficiency.

3) Transformer-based models

Vision Transformers (ViTs) and hybrids have emerged as powerful alternatives. Models such as Swin Transformer [16], SegFormer [17], and TransUNet [18] excel at global context modeling but are computationally demanding and data-hungry, limiting their scalability in aerial segmentation tasks.

4) Ensemble and fusion strategies

Several works have applied decision-level or score-level ensembles, showing improved robustness compared to single networks [12]. However, most prior studies either use homogeneous models (same architecture with different initializations) or computationally expensive feature-level fusion. Few works explore heterogeneous mask-based fusion of U-Net and U-Net++, which is the main contribution of this study.

2.3. Transformer-based segmentation models

Recent advancements in semantic segmentation highlight the role of transformer-based architectures that leverage global receptive fields and long-range dependency modeling. Models such as Swin Transformer [19], SegFormer [17], and hybrid encoder–decoder approaches like TransUNet [18] demonstrate strong performance in remote sensing tasks. However, these methods typically require substantial computational resources and large-scale training datasets, which may limit applicability in low-resource deployment contexts. The proposed DCAMF framework offers a competitive alternative by enhancing classical CNN architectures with confidence-aware ensemble modeling, providing improved performance while maintaining lower architectural complexity.

2.4. Comparative summary

Table 1 provides a comparative overview of the representative studies reviewed in this work, highlighting the employed segmentation models, datasets, evaluation metrics, key findings, and reported limitations. This comparison facilitates the identification of current research trends and underscores the need for a robust, confidence-aware fusion framework that can effectively combine the complementary strengths of heterogeneous segmentation architectures.

2.5. SLR findings and research gap

The systematic review reveals three consistent patterns:

1) CNN baselines offer efficient computation but struggle with fine boundaries and small-object segmentation.
2) U-Net variants achieve better localization, but their performance depends strongly on the chosen backbone.
3) Transformer-based models provide superior global context modeling but require large-scale annotated data and high computational resources.
4) Existing fusion strategies either combine homogeneous models or use score-level fusion, leaving a gap for heterogeneous mask-based fusion of complementary architectures such as U-Net and U-Net++.

The present study addresses this gap by developing a mask-level fusion framework that leverages the strengths of U-Net and U-Net++ with diverse backbones, achieving higher robustness and generalization across multi-class aerial and drone datasets.

Existing ensemble-based segmentation methods typically rely on homogeneous architectures or static weighting schemes. In contrast, the proposed framework explicitly combines heterogeneous encoder–decoder models and introduces dynamic

Table 1
Representative studies on aerial/drone semantic segmentation

Study	Model(s)	Dataset	Classes	Metric(s)	Key findings	Limitation
[20]	U-Net	Aerial imagery	6	PA = 76–78%, mIoU = 0.55	Established baseline with CNN encoder–decoder	Limited accuracy
[21]	Circle-U-Net	Aerial dataset	5	PA = 82%	Improved boundaries via circular skip paths	High computation
[22]	Hybrid CNN	Dubai imagery	5	PA = 82%	Hybridization improved localization	Limited dataset
[14]	DensePlusU-Net	Dubai dataset	6	PA = 86.1%	Dense connections enhanced learning	Backbone heavy
[23]	U-Net	UAV crop images	1	PA = 97%	Strong performance on binary task	Single-class only
[19]	Swin Transformer	Aerial dataset	8	mIoU = 0.67	Excellent global context	Data-hungry
[17]	SegFormer	Aerial imagery	8	mIoU = 0.65	Efficient transformer	Requires large-scale data
[18]	TransUNet	Remote sensing	10	PA = 84%, mIoU = 0.60	Hybrid CNN + ViT effective	Heavy computation
[12]	Fusion ensembles	Multi-modal	6	PA ≈ 85%	Ensembles improve robustness	Score-level fusion only
This work	DCAMF (U-Net + U-Net++)	Drone (23), Aerial (6)	23 + 6	PA = 89.25%, mIoU = 0.671	Robust multi-class segmentation, efficient	Novel approach

confidence weighting at the mask level, which distinguishes it from recent CNN-based ensembles and transformer-driven approaches that require substantial computational overhead.

3. Proposed DCAMF Framework for Aerial Image Segmentation

3.1. Overview and contributions

This section presents the proposed Dynamic Confidence-Aware Mask Fusion (DCAMF) framework for robust semantic segmentation of aerial imagery. The framework is designed to address the challenges of high intra-class variability, complex object boundaries, and scale diversity commonly encountered in aerial and drone-based datasets.

The proposed approach integrates heterogeneous encoder–decoder segmentation architectures, namely, U-Net and U-Net++, combined with multiple backbone networks including MobileNetV2, VGG16, ResNet50, and EfficientNet. Instead of relying on a single segmentation model or static ensemble averaging, DCAMF operates at the mask (decision) level, where the outputs of multiple segmentation models are dynamically fused based on confidence-aware weighting.

The main contributions of this work are summarized as follows:

1) A novel DCAMF framework that unifies U-Net and U-Net++ architectures with diverse backbone encoders for aerial image segmentation.
2) A mathematically grounded fusion mechanism that incorporates pixel-level confidence estimation into the mask fusion process.
3) A transparent and reproducible fusion pipeline supported by algorithmic description, mathematical formulation, and a visual workflow.
4) Extensive experimental validation on multi-class aerial datasets demonstrating consistent improvements over single-model baselines and conventional fusion strategies.

3.2. System architecture and workflow

Figure 1 illustrates the overall architecture of the proposed DCAMF framework. Given an input aerial image, multiple segmentation models based on U-Net and U-Net++ architectures are executed in parallel. Each model employs a different backbone encoder to extract complementary spatial and semantic features.

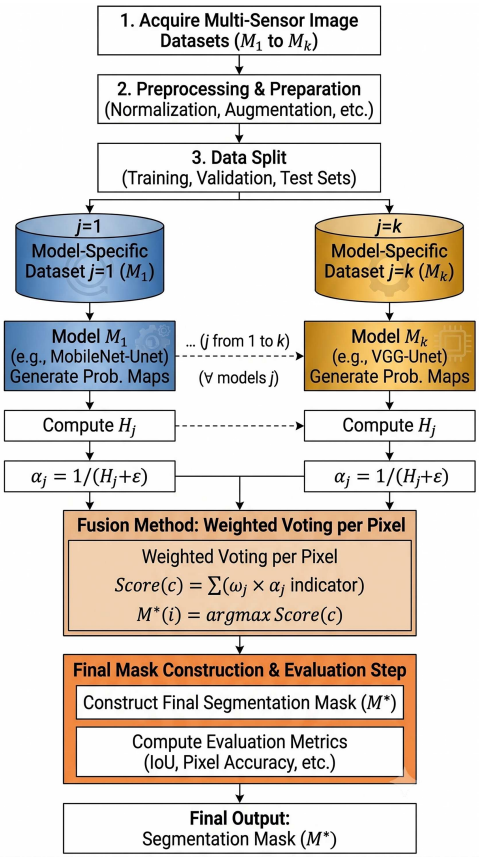
As shown in Figure 1, the fusion process operates at the mask level, ensuring that model diversity is preserved while enabling spatially adaptive decision-making based on pixel-level confidence.

Each base model independently produces a pixel-wise segmentation mask. These masks are subsequently passed to the DCAMF fusion module, which performs confidence-aware mask aggregation to generate a final fused segmentation output. By deferring the fusion operation to the decision level, the framework preserves model diversity while reducing the propagation of shared feature-level errors.

To ensure architectural diversity, the framework employs two complementary encoder–decoder segmentation paradigms: U-Net and U-Net++.

U-Net is characterized by a symmetric encoder–decoder structure with skip connections that facilitate the preservation of spatial information. In contrast, U-Net++ introduces nested skip connections and dense feature aggregation, enabling improved boundary refinement and multi-scale feature fusion.

Figure 1
DCAMF framework integrating U-Net and U-Net++ with diverse backbones



Both architectures are instantiated with multiple backbone encoders, including MobileNetV2, VGG16, ResNet50, and EfficientNet. This design choice promotes feature diversity across models and forms the foundation for effective decision-level fusion.

3.3. Datasets

The proposed framework is evaluated on two publicly available aerial image datasets with different class distributions and complexity levels. The first dataset contains high-resolution drone imagery with fine-grained semantic classes, while the second dataset focuses on aerial scenes with fewer but visually diverse land-cover categories. This dataset diversity enables a comprehensive assessment of the robustness and generalizability of the proposed fusion strategy.

3.4. Dynamic Confidence-Aware Mask Fusion (DCAMF)

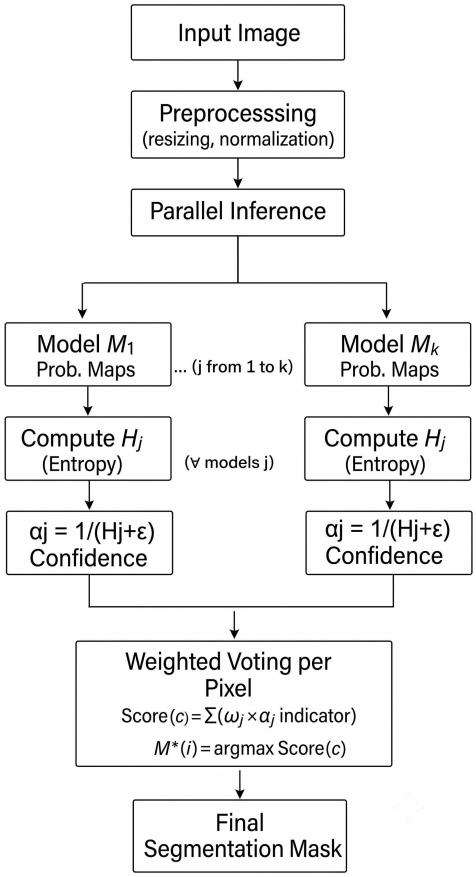
3.4.1. Algorithmic description

The DCAMF module fuses segmentation masks generated by multiple base models using a dynamic, confidence-driven strategy. For each input image, the module receives a set of predicted segmentation masks corresponding to different architectures and backbones.

For each pixel location, DCAMF evaluates the confidence of each model's prediction and assigns an adaptive weight accordingly. The final class label is determined through a weighted aggregation process, ensuring that predictions from more confident models exert greater influence. This dynamic behavior allows the fusion mechanism to adapt spatially across the image rather than relying on static global weights.

Figure 2 illustrates the steps of parallel model inference, entropy-based confidence estimation, adaptive weighting, and pixel-wise weighted voting to generate the final fused mask.

Figure 2
Workflow of the proposed DCAMF fusion framework



To clearly describe how the final segmentation mask is generated, the proposed fusion process is summarized through an explicit algorithmic workflow and a visual flowchart. This clarification details the steps used to combine multiple segmentation masks into a single fused output.

In order to enhance the robustness and address the limitations of static weighting, we extend the fusion operator by integrating model confidence through an entropy-based reliability score. The proposed mechanism is termed Dynamic Confidence-Aware Mask Fusion (DCAMF) and estimates adaptive pixel-level contribution for each model during fusion. To address the lack of adaptiveness in existing ensemble segmentation approaches, we introduce a novel decision-level fusion mechanism named DCAMF, which incorporates model-specific confidence through entropy-based weighting. This enhancement enables the fused output to dynamically prioritize predictions with higher reliability.

The proposed DCAMF fusion process operates through five stages: preprocessing, prediction, entropy-based confidence evaluation, adaptive weight computation, and final mask aggregation. Algorithm 1 summarizes the overall procedure.

Algorithm 1: DCAMF-Based Mask Fusion

Input: Set of models M1. . . Mk, input image I
Output: Final fused mask M*
1: For each model Mj:
2: Predict segmentation mask → Mj(I)
3: Compute class probabilities and entropy Hj
4: Convert entropy to confidence αj = 1 / (Hj + ε)
5: Normalize αj across all models
6: For each pixel index i:
7: For each class c:
8: Score(c) = Σj (ωj × αj × I(Mj(i)=c))
9: M*(i) = argmax_c Score(c)
10: return M*

3.4.2. Mathematical formulation

1) Model architectures

We employed U-Net and U-Net++ as the base architectures. Four different backbone encoders were integrated: MobileNetV2, VGG16 [24], ResNet50 [25], and EfficientNet. Each backbone was initialized with ImageNet pre-trained weights and fine-tuned on aerial datasets.

a. **U-Net** employs a symmetric encoder–decoder design with skip connections:

$$F_{U-Net}(X) = S(X) + D(E(X)) \tag{1}$$

where $E(X)$ is the encoder feature extraction, $D(\cdot)$ is the decoder upsampling, and $S(X)$ represents skip connections.

b. **U-Net++** introduces dense nested skip pathways:

$$F_{++U-Net}(X) = D\left(E(X), \{Sij(x)\}_{ij}\right) \tag{2}$$

where $\{Si,j(X)\}$ denotes hierarchical skip connections indexed by encoder depth i and pathway j. This improves multi-scale feature fusion and boundary localization.

2) Mask-based fusion strategy

Let M = $\mathcal{M}$ {M1,M2,. . . ,Mk} denote the set of segmentation masks generated by k models (U-Net and U-Net++ variants with different backbones). Each mask assigns a predicted label $i(j)$ y for pixel i in mask Mj.

We define the fusion operator F at the decision (mask) level as:

$$M^*(i) = \left(c = i^{(j)}y\right) 1.arg\max_{c\in C} \sum_{j=1}^{k} w_j \tag{3}$$

Where:

- $M*(i)$ is the fused label for pixel i.
- C is the set of classes.
- ω_j is the weight assigned to model j, calibrated from validation mIoU.
- $1(\cdot)$ is the indicator function (1 if true, else 0).

This formulation corresponds to weighted majority voting at the pixel level. If $\omega_j = w_j$ k/1, it reduces to unweighted majority voting.

The weights are adaptively computed as:

$$w_j = \frac{mIoU_j}{k_{i=1}\sum mIoU_i} \tag{4}$$

Ensuring that models with higher validation performance contribute more strongly to the final decision.

To address the limitations of conventional ensemble strategies that rely on uniform or majority voting, this work introduces a mathematically grounded, uncertainty-aware mask fusion framework. The proposed formulation enables pixel-wise adaptive decision-making by explicitly modeling both prediction confidence and inter-model reliability. In the following, a step-by-step mathematical formulation of the proposed fusion mechanism is presented.

3) Probabilistic output modeling

Let $P_j(i)$ denote the class probability distribution predicted by the j-th segmentation model at pixel i. This distribution is defined as:

$$P_j(i) = \{C \ni c | p_j(i,c)\} \tag{5}$$

where C represents the set of semantic classes and $p_j(i, c)$ denotes the probability that pixel i belongs to class c according to model j.

4) Pixel-wise mask generation

Based on the predicted probability distribution, each model generates a segmentation mask by assigning to each pixel the class with the maximum posterior probability:

$$(i)_{\hat{j}}\, y = arg\max_{c\in C} p_j(i,c) \tag{6}$$

This operation produces an individual segmentation mask for each model, which serves as the basis for subsequent fusion.

5) Uncertainty quantification via entropy

To quantify the prediction uncertainty at each pixel, Shannon entropy is employed as follows:

$$H_j(i) = \sum\nolimits_{c\in C} - p_j(i,c) \log\left(p_j(i,c)\right) \tag{7}$$

A higher entropy value indicates greater uncertainty in the model's prediction, while lower values reflect higher confidence.

6) Confidence-aware weight assignment

Based on the estimated uncertainty, a confidence-aware weight is assigned to each model at the pixel level:

$$\alpha_j(i) = \frac{1}{H_j(i) + \epsilon} \tag{8}$$

where ε is a small constant introduced to ensure numerical stability. This formulation ensures that predictions with lower uncertainty contribute more strongly to the fusion process.

7) Final mask fusion rule (core contribution)

The final semantic label for pixel i is obtained by aggregating the weighted predictions of all models as follows:

$$(i)\hat{y} = (c = arg\max_{c\in C} \sum_{j=1}^{K} w_j\, \alpha_j(i) \prod(\hat{y}_j(i) \tag{9}$$

where $_j\omega$ denotes the global reliability weight of model j and $(\cdot)$ II is the indicator function. Unlike conventional ensemble methods, the proposed fusion strategy performs pixel-wise, uncertainty-aware decision-making, enabling robust segmentation in regions characterized by high intra-class variability or ambiguous boundaries.

The key scientific contribution of this work lies in formulating mask fusion as an uncertainty-aware optimization problem rather than a heuristic ensemble. By integrating entropy-based confidence modeling with pixel-level weighted decision rules, the proposed framework introduces a principled and extensible fusion strategy that enhances robustness without requiring additional training or architectural modifications.

3.5. Evaluation metrics

The segmentation performance of the proposed framework is evaluated using standard metrics, including pixel-wise accuracy, mean Intersection over Union (mIoU), and validation loss. These metrics provide complementary insights into classification correctness, spatial overlap quality, and convergence behavior, respectively.

4. Results and Discussion

4.1. Experimental setup

This section presents the experimental results of the proposed DCAMF framework. Before detailing the $\hat{j}^{y}$ quantitative evaluations, it is important to clarify the design rationale of DCAMF and its expected impact on segmentation performance.

As established in Section 3, DCAMF is intentionally designed as a decision-level, confidence-aware fusion framework rather than a feature-level or score-level ensemble. This choice is motivated by the observation that different segmentation architectures and backbone encoders exhibit heterogeneous strengths across spatial regions, object scales, and semantic classes in aerial imagery. Consequently, enforcing a uniform fusion strategy or relying on a single dominant model often leads to suboptimal generalization.

By operating at the mask level, DCAMF preserves architectural diversity while mitigating correlated errors introduced during feature extraction. The confidence-driven weighting mechanism enables spatially adaptive fusion, allowing more reliable models to dominate the decision in regions where they demonstrate higher predictive certainty. This dynamic behavior is particularly beneficial in aerial datasets characterized by high intra-class variability, complex boundaries, and class imbalance.

The experimental design in Section 4 directly reflects this theoretical foundation. The evaluated scenarios systematically analyze (1) individual U-Net and U-Net++ models, (2) homogeneous DCAMF fusion within each architecture family, and (3) heterogeneous DCAMF fusion across U-Net and U-Net++. This progressive evaluation strategy enables a clear assessment of the contribution of each design component and provides empirical evidence supporting the superiority of the proposed DCAMF framework.

To leverage prior knowledge and accelerate convergence, all backbone networks (MobileNetV2, VGG16, ResNet50, EfficientNet) were initialized with ImageNet pre-trained weights and subsequently fine-tuned on the target datasets. The training process employed the Adam optimizer with an initial learning rate of $\times 14{-}10$, dynamically reduced upon plateau in validation loss. A mini-batch size of 8 was selected to balance computational efficiency with gradient stability, while early stopping with a patience of 15 epochs was applied to prevent overfitting. Training was conducted for up to 200 epochs, depending on convergence behavior.

To improve model robustness and mitigate dataset imbalance, extensive data augmentation was applied during training. This included geometric transformations (random rotations, horizontal and vertical flips, scaling, and cropping), photometric enhancements (contrast stretching, brightness and gamma correction), and spatial adjustments (resizing and padding). Such augmentations not only increased the effective dataset size but also simulated real-world variations in illumination, scale, and viewing angle, making the models more resilient to environmental diversity.

In addition, the dataset was partitioned into 70% training, 15% validation, and 15% testing to ensure fair evaluation. Validation sets guided hyperparameter tuning and model selection, while test sets remained unseen until the final performance assessment. All reported results therefore reflect generalization ability rather than memorization, highlighting the reliability of the proposed fusion framework.

4.2. Training scenarios

To comprehensively evaluate the effectiveness of the proposed fusion framework, a series of structured training and evaluation scenarios were designed. These scenarios were applied to two benchmark datasets of varying complexity: the Semantic Drone Dataset (23 classes) and the Aerial Imagery Dataset (6 classes). By systematically analyzing model behavior across datasets and configurations, the experiments aimed to reveal the strengths and limitations of both individual architectures and fusion strategies.

The training scenarios are summarized as follows:

1) Baseline U-Net models with multiple backbones

The U-Net architecture was trained independently on each dataset using four distinct backbone encoders: **MobileNetV2, ResNet50, EfficientNet, and VGG16**. This step established baseline performance and allowed a comparative analysis of backbone contributions in terms of segmentation accuracy and computational efficiency.

2) DCAMF (U-Net backbones)

To exploit the complementary strengths of different encoders, the top-performing U-Net models were fused at the mask level. This evaluated whether decision-level fusion among backbones could reduce error variance and enhance segmentation robustness.

3) Baseline U-Net++ models

Based on the backbone performance observed in the U-Net experiments, the most promising backbones were integrated into the **U-Net++ architecture**. The nested skip connections in U-Net++ were expected to further improve feature aggregation and boundary precision.

4) DCAMF (U-Net++ backbones)

Similar to U-Net, decision-level fusion was applied among U-Net++ models using different backbones. This scenario assessed whether nested architectures benefited equally from fusion strategies.

5) DCAMF (U-Net + U-Net++)

A novel experiment combined masks generated by **U-Net and U-Net++ models**. This heterogeneous fusion exploited the diversity between skip connections in U-Net and dense pathways in U-Net++, targeting further improvements in class-wise mean Intersection over Union (mIoU) and overall generalization.

6) Repetition across datasets

All the above scenarios (1–5) were repeated on the second dataset (Aerial Imagery with 6 classes). This cross-dataset evaluation ensured that the conclusions drawn were not dataset-specific but instead reflected the **generalizability of the fusion approach** across different data distributions and class granularities.

4.2.1. Preprocessing operations

The specific values [0.485, 0.456, 0.406] for the mean and the standard deviation values [0.229, 0.224, 0.225] were derived from the ImageNet dataset, which is commonly used for pre-training deep learning models. All images are transformed into RGB format. All training images are loaded along with the corresponding masks. The data augmentations and albuminations that are applied to the training and validation datasets are:

1) Resizing: The images are resized into a fixed size [704, 10556] for the drone dataset and 256 × 256 for the aerial image dataset using the nearest neighbor interpolation method.
2) Horizontal and vertical flips.
3) Grid distortion with a percentage of 0.2.
4) Random contrast in the range (0–0.5)

4.2.2. Training parameters

To ensure robust training and fair evaluation, dataset-specific partitioning and carefully tuned hyperparameters were employed. For the Semantic Drone Dataset (23 classes), the data were split into 75% training (306 images), 15% validation (54 images), and 10% testing (40 images). This ensured sufficient training data while retaining enough samples for reliable validation and unbiased testing. For the Aerial Imagery Dataset (6 classes), the data were divided into 80% training (58 images), 20% validation, and 14 images for testing. Given the smaller dataset size, a slightly larger validation portion was reserved to monitor generalization and mitigate overfitting.

The training relied on hyperparameters optimized through extensive preliminary trials and iterative fine-tuning. Multiple experiments were conducted to evaluate the impact of varying batch size, learning rate, optimizer configurations, and regularization. The final adopted configuration is summarized in Table 2, reflecting the most effective trade-off between convergence speed, stability, and segmentation accuracy.

These parameters were not arbitrarily chosen; rather, they emerged as the optimal configuration after extensive trial and error and guided by prior literature on deep learning for semantic segmentation. Early stopping and weight decay further enhanced generalization, ensuring that the models did not overfit to training data.

This methodological tuning process demonstrates the rigor and depth of the experimental design, confirming that the reported results are the outcome of systematic optimization rather than arbitrary parameter selection.

4.3. Results of training scenarios for the first dataset

The results obtained on the Semantic Drone Dataset (23 classes) are presented according to the designed training scenarios. For each experiment, we tracked loss per epoch, pixel accuracy, and mIoU on both training and validation sets. These metrics were critical to evaluate convergence stability, overfitting behavior, and segmentation performance.

4.3.1. Results of the U-Net architecture and DCAMF (all U-Net backbones)

The baseline experiments involved training the U-Net architecture with four different backbones: MobileNetV2, ResNet50, EfficientNet, and VGG16. To ensure reliability, training was repeated multiple times, with early stopping applied whenever validation loss failed to decrease consistently.

As an illustrative case, Figure 3 shows the training curves of U-Net with the MobileNetV2 backbone. The training process was executed 27 times before early stopping was triggered, demonstrating the stability of the learning procedure. The final performance metrics achieved were:

1) Validation accuracy: 88.8%
2) Validation mIoU: 0.435
3) Validation loss: 0.371

In the fusion step, the DCAMF was first applied to the three best-performing U-Net models, followed by the fusion of all U-Net backbones. This hierarchical fusion analysis provided insight into whether performance gains arise from selective combinations (top-3 models) or from comprehensive aggregation of all backbones, as shown in Figure 3.

Table 2
Training hyperparameters and experimental settings

Parameters	Rate/condition
Batch size	5
Learning rate	1e-3
Epochs	30
Weight decay	1e-4
Loss function	Categorical cross-entropy
Optimizer	Adam
Saving model	Save the model only if the loss is decreased
Early stop condition	Depending on the validation loss, the training is terminated if it has not decreased seven times during the training process
Initialization	ImageNet pre-trained weights
Hardware	GPU 15GB VRAM/12.7GB RAM
Frameworks	TensorFlow – Keras – PyTorch

Figure 3
The loss, mIoU, and accuracy of training U-Net with MobileNetV2 on a drone dataset

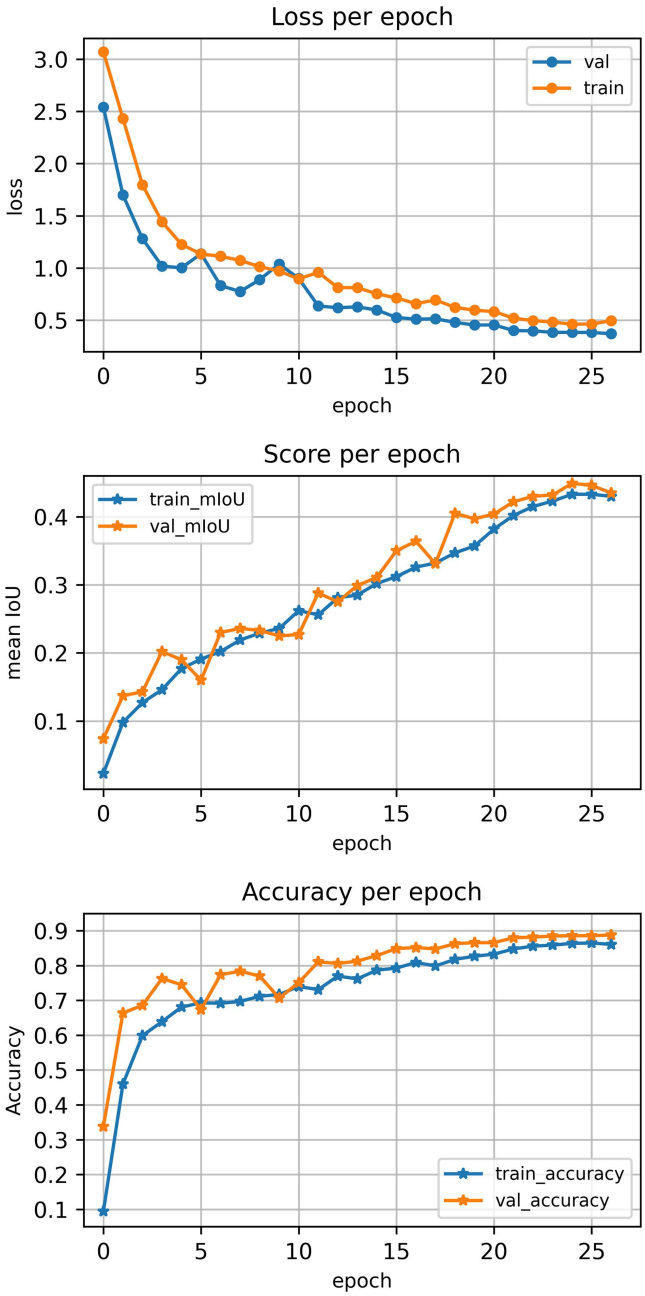


Figure 4
The loss, mIoU, and accuracy of training U-Net with VGG16 on a drone dataset

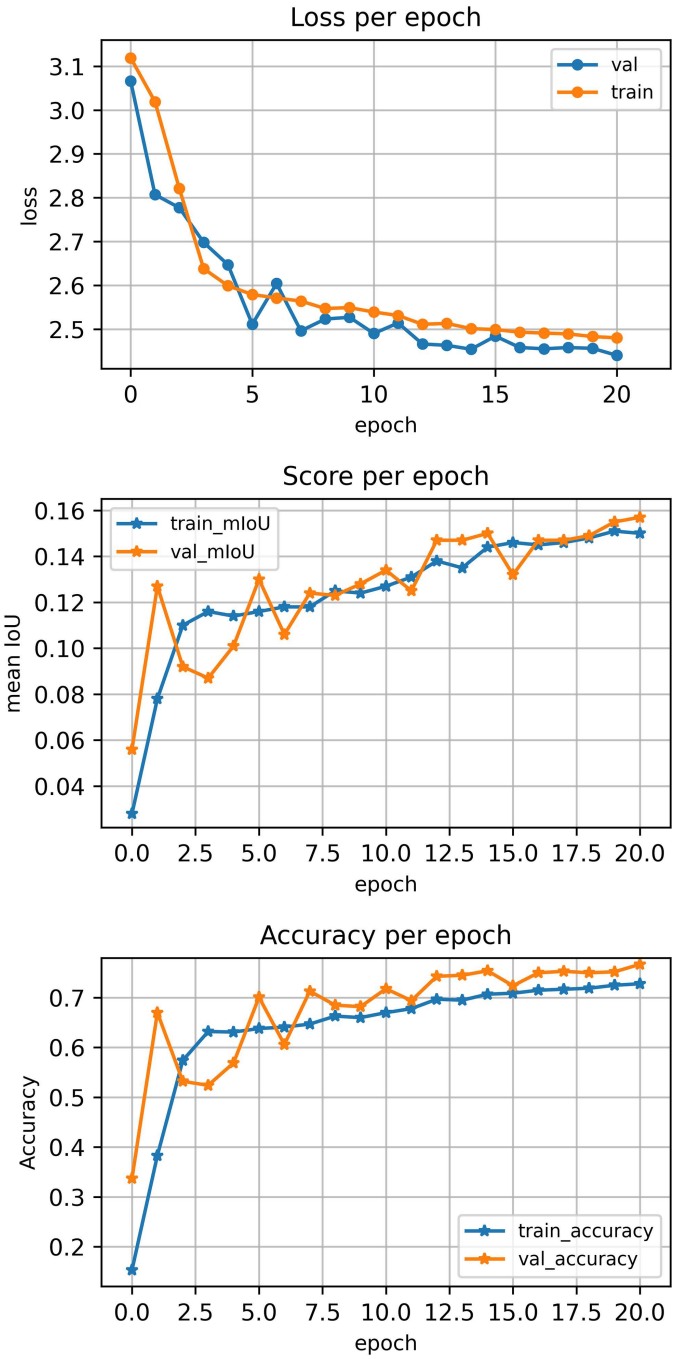


Figure 4 presents the training results of U-Net with VGG16 backbone. The training process stopped early after 21 repetitions, yielding a validation accuracy of 76.7%, mIoU of 0.157, and a validation loss of 2.44. Compared to MobileNetV2, the performance of VGG16 was significantly lower, confirming that this backbone is not suitable for aerial semantic segmentation in the given setting.

Figure 5 shows the training outcomes of U-Net with EfficientNet backbone. The training stopped early after 24 repetitions, achieving a validation accuracy of 80.9%, mIoU of 0.203, and a validation loss of 2.399. Although better than VGG16, the results remain inferior to MobileNetV2, indicating limited effectiveness of EfficientNet for this dataset.

Figure 6 presents that the U-Net model with ResNet50 backbone was trained 19 times, achieving a validation accuracy of 80.8%, mIoU of 0.3, and validation loss of 0.595. While slightly lower than MobileNetV2, ResNet50 outperformed other tested models, demonstrating strong feature extraction and stable performance across repeated training cycles.

Figure 5
The loss, mIoU, and accuracy of training U-Net with EfficientNet on a drone dataset

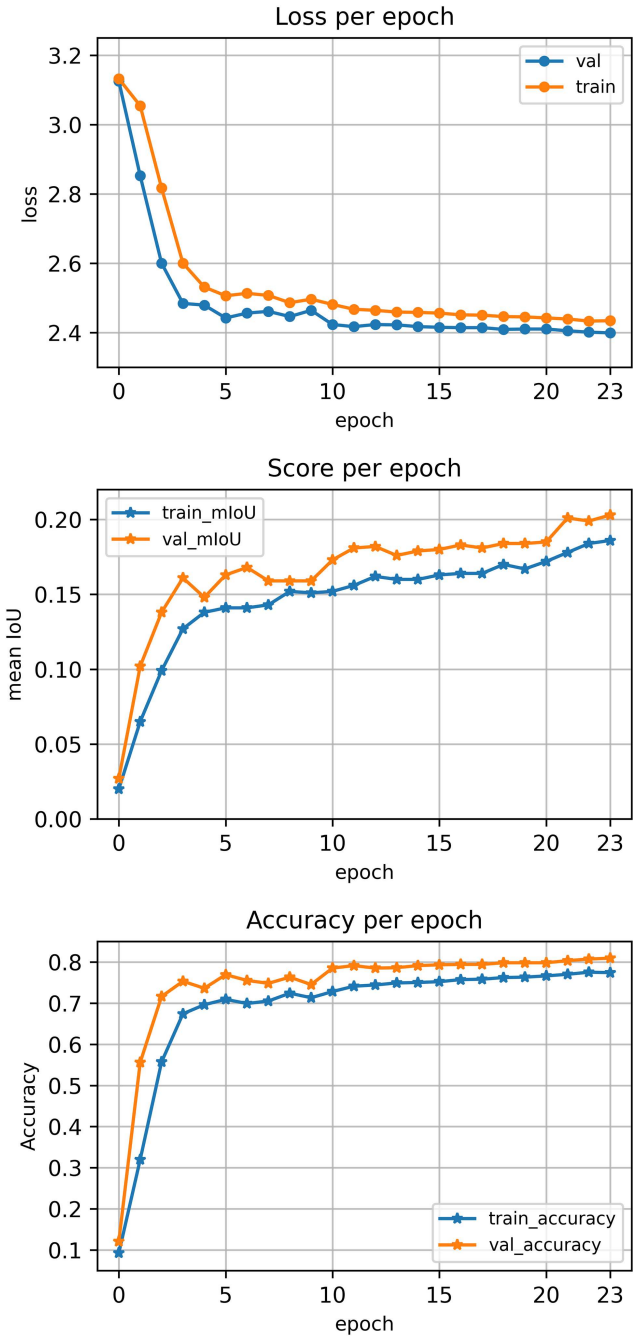


Figure 6
The loss, mIoU, and accuracy of training U-Net with ResNet50 on a drone dataset

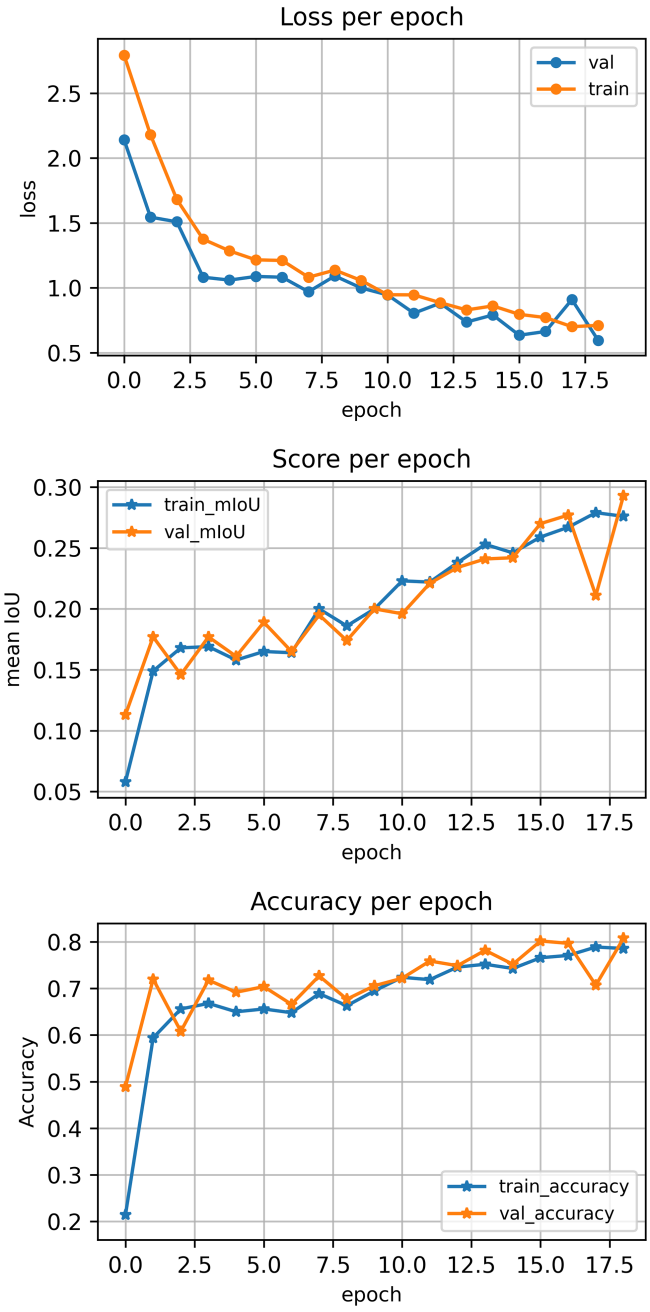


Figure 7 presents sample segmentation results on the test set using U-Net with MobileNetV2 backbone, illustrating accurate delineation of target regions and the model's effectiveness in capturing fine-grained spatial features.

Table 3 summarizes the performance comparison between individual U-Net models and the fusion model, highlighting the improvements achieved through model integration.

To validate the effectiveness of the proposed DCAMF mechanism, we compared its performance against standard ensemble techniques including Majority Voting, Averaging, and Static Weighted Voting. As shown in Table 3, DCAMF consistently improved pixel accuracy and mIoU across both datasets, confirming its effectiveness in reducing uncertainty propagation from low-confidence model predictions.

Table 3 demonstrates that the proposed DCAMF model outperforms individual U-Net models, achieving a 0.567% higher test accuracy and a 0.05 (5%) increase in mIoU compared to the best individual model, MobileNetV2. Figure 8 illustrates another example demonstrating the advantages of the fusion model.

The results in Table 3 and Figure 8 confirm that the DCAMF model achieves the highest accuracy and significantly enhances performance.

Figure 7
Some examples of the segmentation results of the MobileNetV2-based U-Net architecture

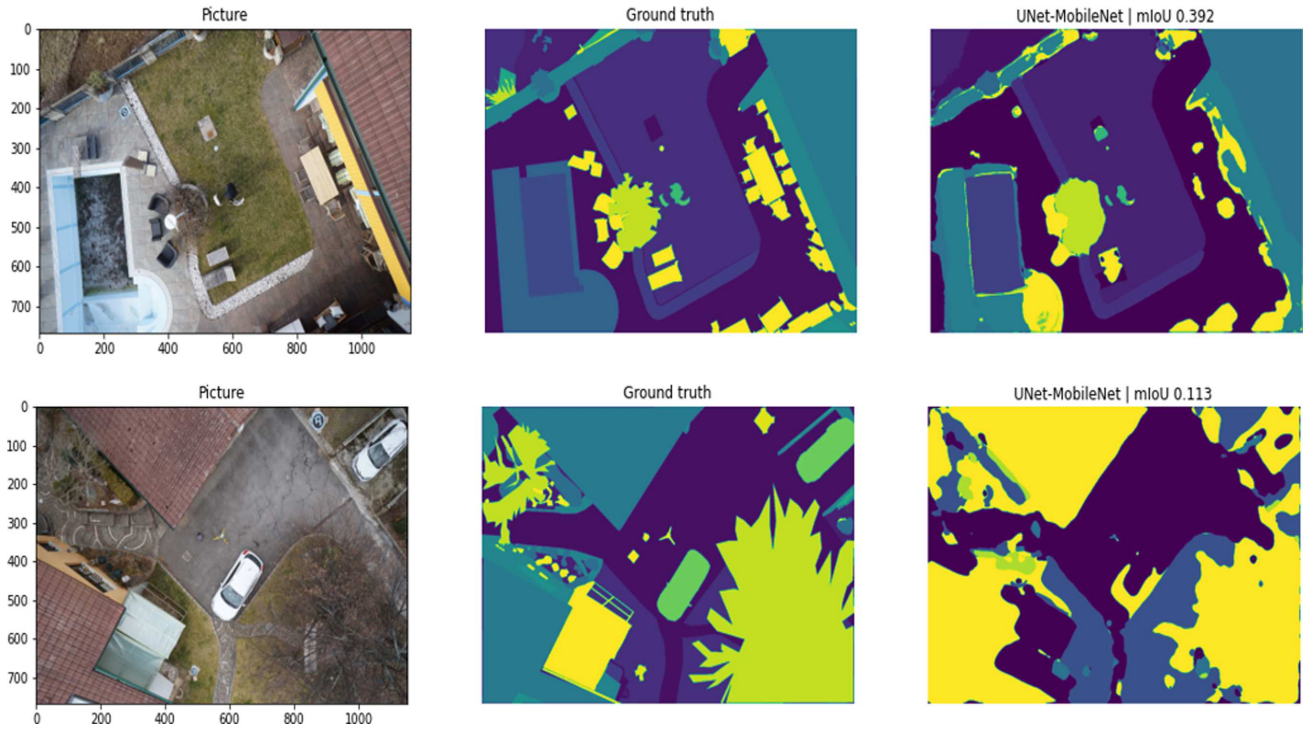
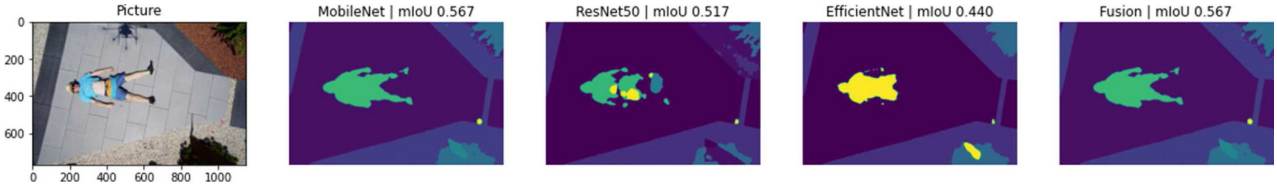


Table 3
Comparison between individual U-Net models and fusion models

Model	Test accuracy (%)	Test mIoU
U-Net MobileNetV2	88.69	0.448
U-Net VGG16	73.41	0.193
U-Net EfficientNet	80.63	0.3246
U-Net ResNet50	80.41	0.257
DCAMF (Top-3 U-NetBackbones)	89.25	0.499
DCAMF (U-Net)	89.14	0.567

Figure 8
An example of the segmentation results of the best three individual models and the fusion models



4.3.2. Results of the U-Net++ architecture and DCAMF (all U-Net++ backbones)

For U-Net++, MobileNetV2 and EfficientNetV2 backbones were selected based on their top performance in U-Net. Figures 9 and 10 present accuracy, mIoU, and loss curves. Table 4 compares individual U-Net++ models with the DCAMF model.

Table 4 indicates that the DCAMF (All U-Net++ Backbones) offers slight improvements over individual models. Figures 9 and 10 show segmentation examples of the DCAMF U-Net++ model on the drone dataset.

Figure 11 shows includes some examples of using U-Net++ on the drone dataset.

Figure 12 shows that U-Net++ segmentation results are considerably worse than those of the U-Net architecture.

4.3.3. DCAMF (U-Net + U-Net++)

The final training scenario on the drone dataset combines all U-Net and U-Net++ models. Table 5 presents the results of this combined fusion.

Table 4
Comparison between individual U-Net++ models (MobileNetV2, EfficientNet) and the DCAMF model

Model	Test accuracy (%)	Test mIoU
U-Net++ MobileNetV2	75.139	0.224
U-Net++ EfficientNet	74.62	0.217
DCAMF (All U-Net++ Backbones)	75.21	0.272

Figure 9
The loss, mIoU, and accuracy of training U-Net++ with MobileNetV2 on a drone dataset

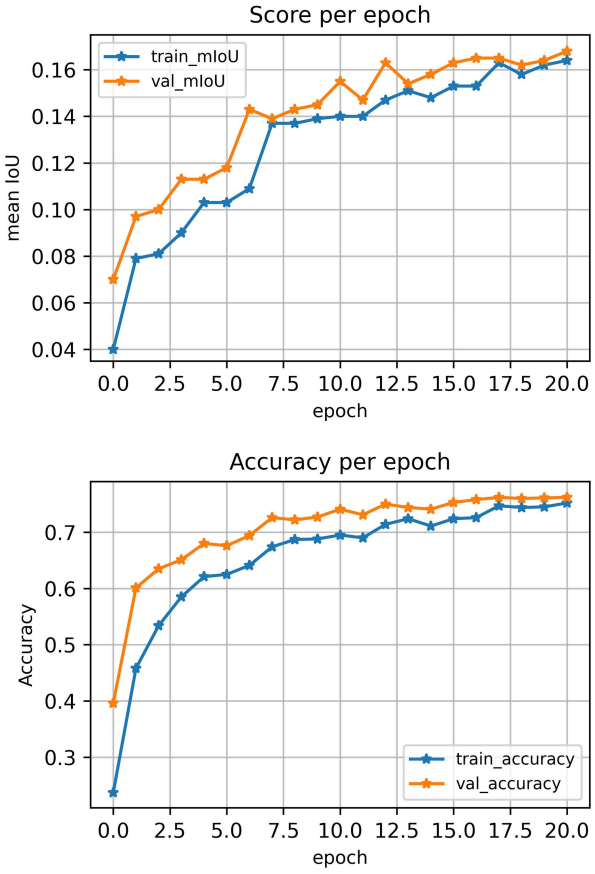


Table 5 indicates that the DCAMF (U-Net Backbones) achieves 89.14% accuracy and an mIoU of 0.5.

4.4. Results of training scenarios for the second dataset

The same fusion and individual model scenarios from the drone dataset were applied to the second aerial dataset.

4.4.1. Results of the U-Net architecture and the DCAMF (all U-Net backbones) on the second dataset

For each scenario, loss, mIoU, and accuracy per epoch were computed for training and validation sets. The fusion step first combined the top three models and then all U-Net models. Table 5 shows that the proposed DCAMF model outperformed individual models. Figure 13 presents U-Net with MobileNetV2, trained 25 times, achieving 87.2% validation accuracy, 0.625 mIoU, and 1.171 validation loss.

Figure 10
The loss, mIoU, and accuracy of training U-Net++ with EfficientNet on a drone dataset

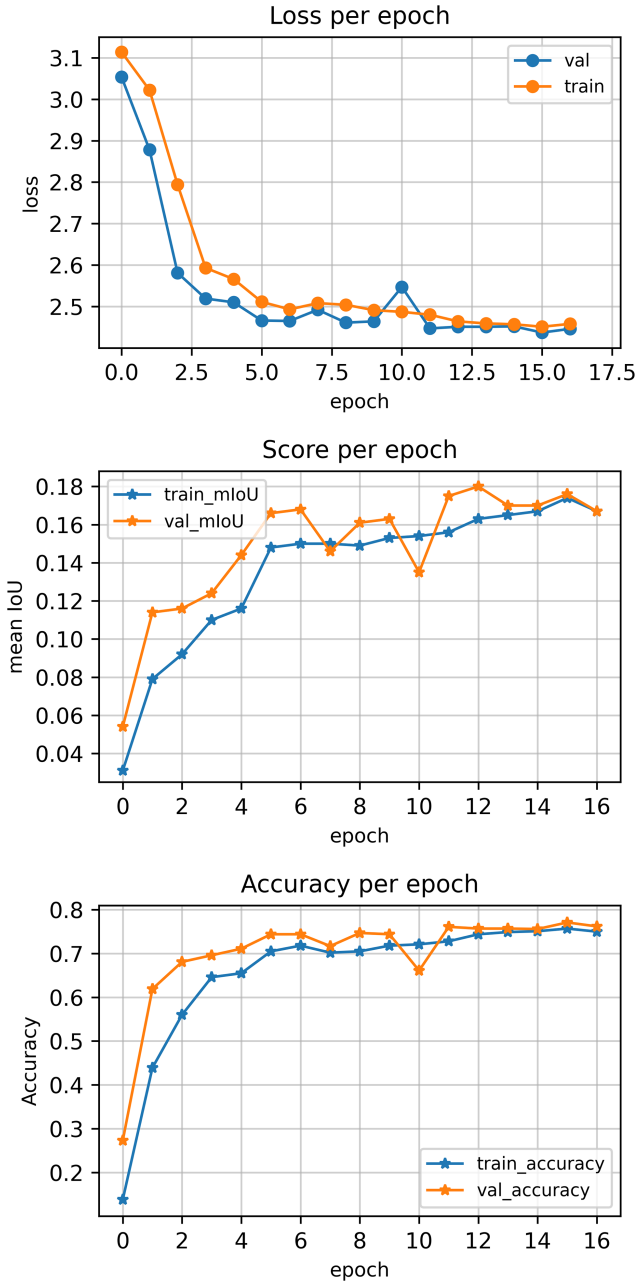


Figure 14 shows U-Net with VGG16 backbone, trained 20 times (early stopping before 10 epochs). Validation accuracy is 85.8%, mIoU 0.604, and loss 1.184. These results improve over the first drone dataset but remain lower than MobileNetV2 performance.

Figure 11
Some examples of segmentation results using the U-Net++ model on the drone dataset

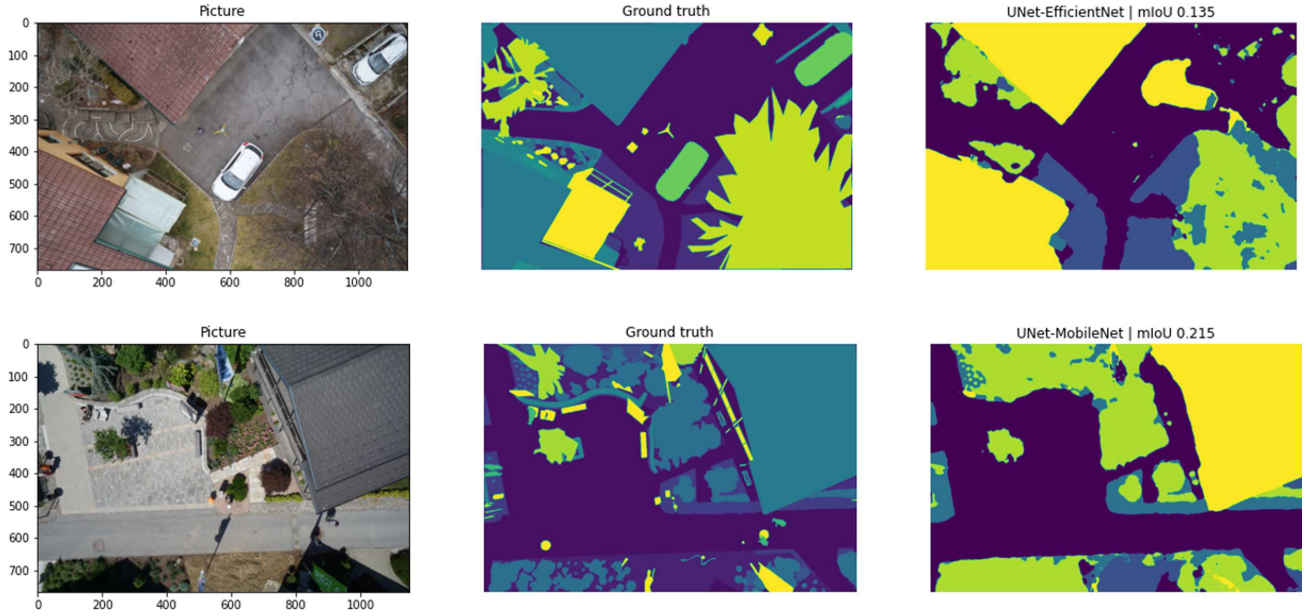


Figure 12
Example of DCAMF (all U-Net++ backbones) on the drone dataset

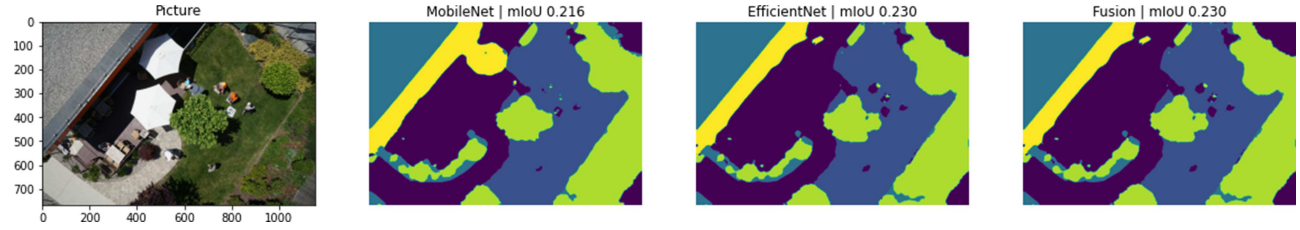


Table 5
Comparison of the DCAMF of U-Net, U-Net++, U-Net, and U-Net++ models

Model	Test accuracy (%)	Test mIoU
DCAMF (U-Net Backbones)	89.14	0.50
DCAMF (U-Net++ Backbones)	75.21	0.272
DCAMF (U-Net + U-Net++)	89.03	0.5143

Figure 15 presents U-Net with EfficientNet backbone, trained 19 times (early stopping before 11 epochs), achieving 87.4% validation accuracy, 0.616 mIoU, and 1.169 validation loss.

Figure 16 shows U-Net with ResNet50 backbone, trained 14 times, achieving 82.6% validation accuracy, 0.551 mIoU, and 1.216 loss, which is lower than both MobileNetV2 and EfficientNet results.

Figure 17 presents test set segmentation results using U-Net across all backbones.

Table 6 compares individual U-Net models (MobileNetV2, VGG16, EfficientNet, ResNet50) with the DCAMF model on the second aerial dataset.

Table 6 shows MobileNetV2 as the best individual backbone, slightly outperforming EfficientNet by 15%. MobileNetV2, VGG16, and EfficientNet all achieved strong performance. The proposed DCAMF model further improves results, with test accuracy up by 0.8% and mIoU increased by 0.0247 (2.47%) over the best individual model.

Figure 18 shows examples of individual and DCAMF models on samples of the test dataset.

The results confirm that the DCAMF model achieves the highest accuracy and significantly enhances performance.

4.4.2. Results of the U-Net++ architecture on the second aerial dataset

For U-Net++, MobileNetV2 and EfficientNetV2 backbones were used, and all U-Net++ models were fused.

Figure 13
The loss, mIoU, and accuracy of training U-Net with MobileNetV2 on the second aerial dataset

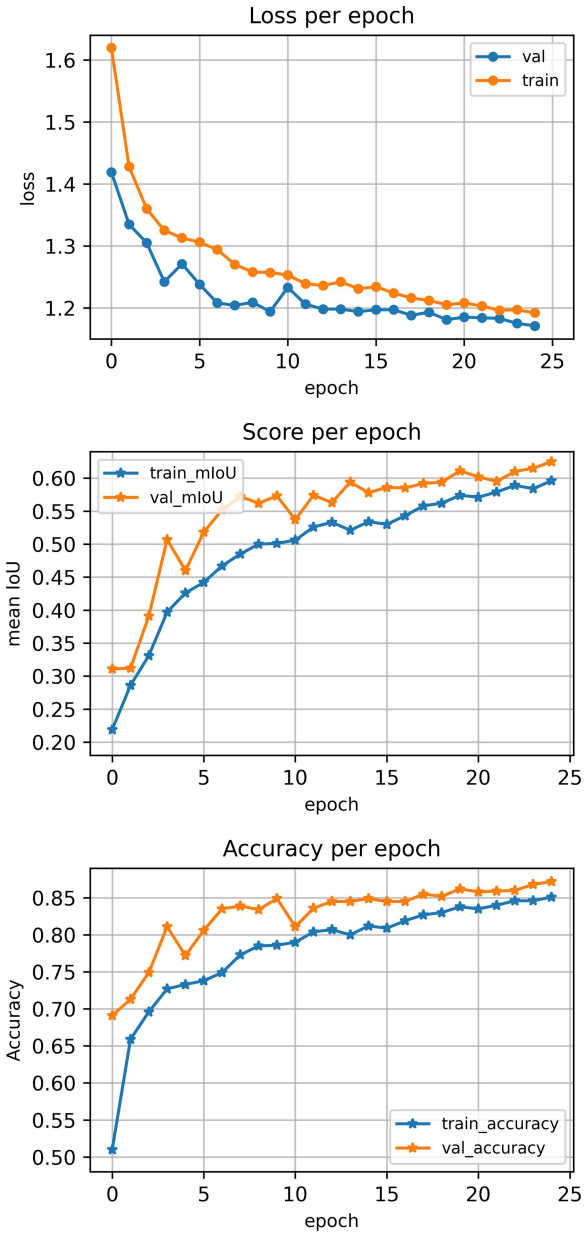


Figure 14
The loss, mIoU, and accuracy of training U-Net with VGG16 on the second aerial dataset

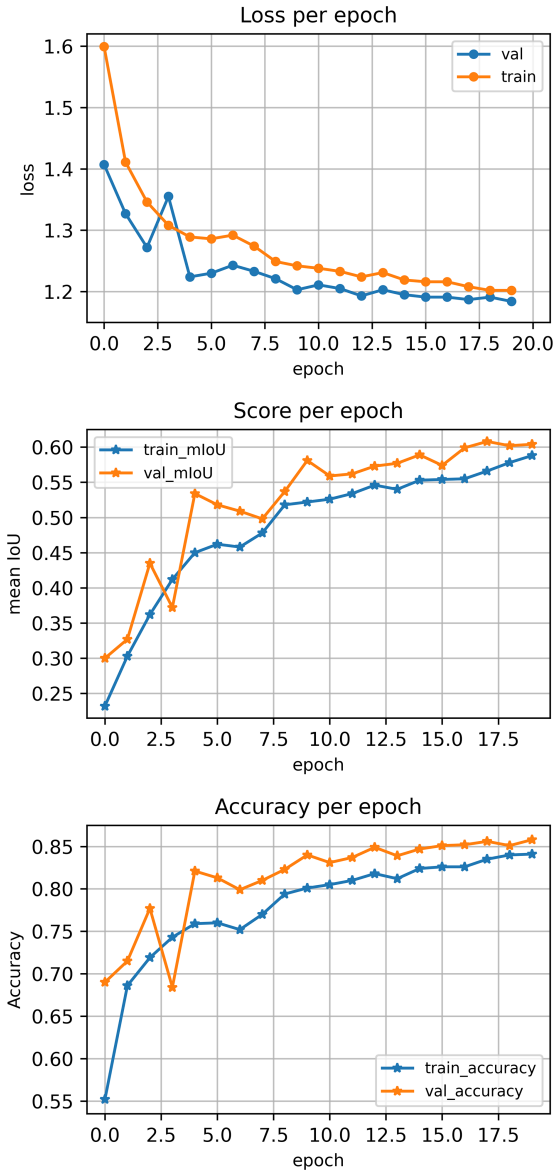


Table 7 compares individual U-Net++ models with the DCAMF model.

Table 7 shows that U-Net++ models have lower accuracy and mIoU than U-Net models. Figures 19 and 20 present accuracy, mIoU, and loss curves. DCAMF of all U-Net++ models provides only slight improvements over individual models.

Figure 21 includes some examples of using U-Net++ on the second aerial dataset.

Figure 21 shows that U-Net++ segmentation results differ on the second aerial dataset. EfficientNet outperforms MobileNet, reversing the trend seen in U-Net and the first drone dataset. This change is attributed to dataset differences: the second dataset has six classes, while the first contains 23.

4.4.3. Results of DCAMF (U-Net + U-Net++) on the second dataset

The final training scenario fuses all U-Net and U-Net++ models. Table 8 presents the results on the second aerial dataset.

Table 8 shows that fusing U-Net and U-Net++ models yields the best results, improving mIoU by 0.0163 (1.63%). U-Net++ models performed well on the second dataset, and their fusion with U-Net further enhances performance. Figure 22 illustrates segmentation results, confirming the performance improvement from this fusion.

4.5. Comparison with related work

Table 9 demonstrates that the proposed DCAMF outperforms existing methods in mIoU and segmentation accuracy. It achieved pixel accuracies of 89.25% and 88.08%, and mIoU

Figure 15
The loss, mIoU, and accuracy of training U-Net with EfficientNet on the second aerial dataset

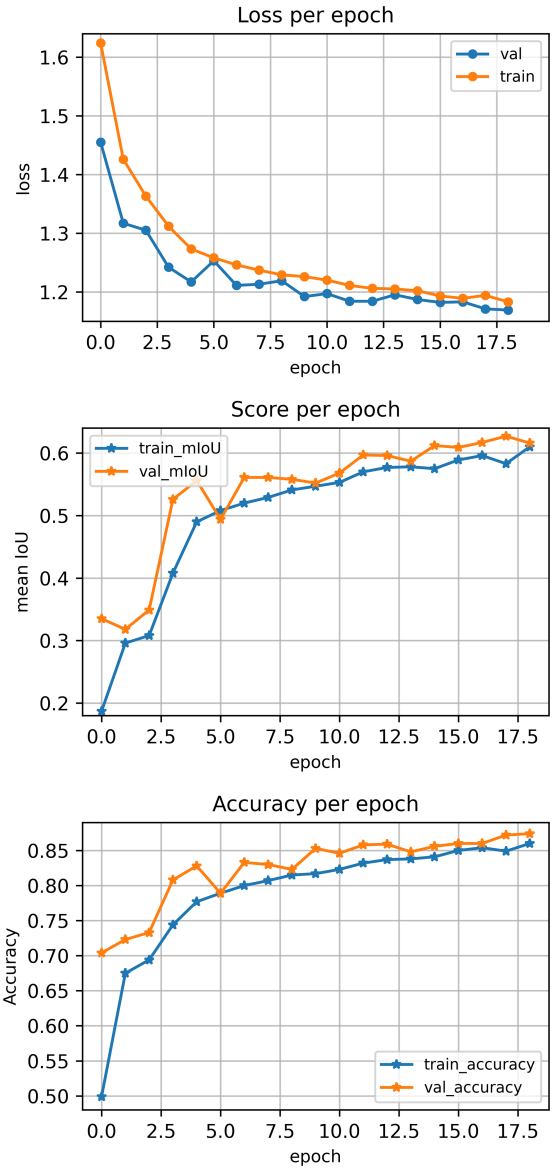
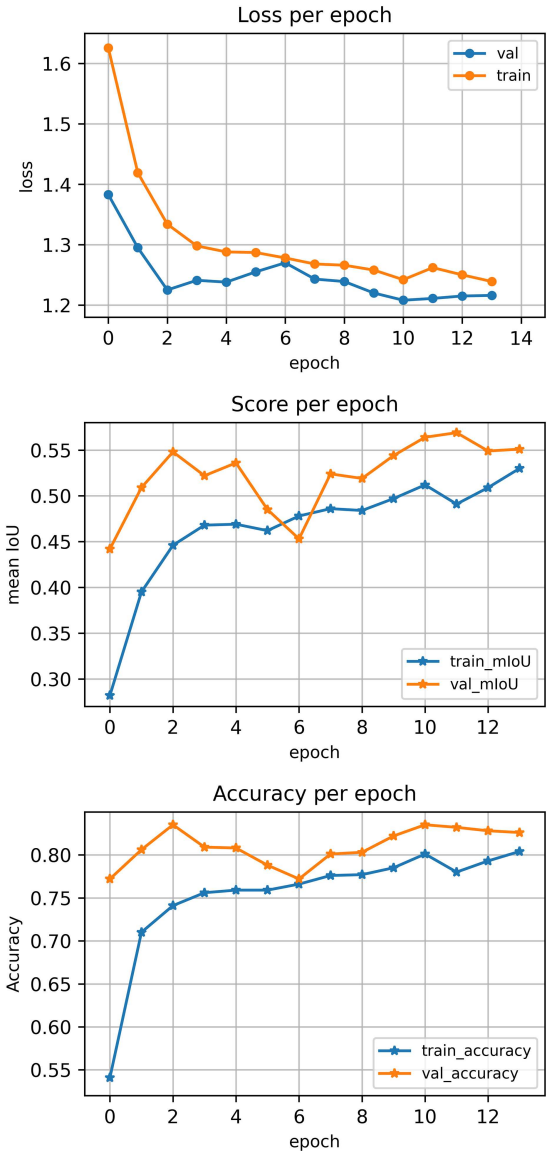


Figure 16
The loss, mIoU, and accuracy of training U-Net with ResNet50 on a drone dataset



values of 0.5143 and 0.671 for the first and second datasets, respectively. The use of fusion techniques contributed to enhanced performance, robustness, and adaptability across datasets. Previous methods, however, faced limitations such as lower accuracy, longer computation times, and missing segmentation metrics.

4.6. Analysis of results

The results reveal several key insights:

DCAMF performance superiority: Fusion consistently outperforms individual models, yielding higher accuracy and mIoU across both datasets, demonstrating effective exploitation of model error diversity.

Dataset-specific fusion: On the Semantic Drone Dataset, fusing three U-Net variants (MobileNetV2, EfficientNet, ResNet50) achieved 89.25% pixel accuracy. On the Aerial Imagery Dataset, including U-Net++ (VGG16) increased accuracy to 88.08%, showing that optimal fusion depends on dataset characteristics.

Enhanced boundary segmentation: Combining U-Net and U-Net++ yielded the highest mIoU (0.5143 for drone, 0.671 for aerial), indicating that U-Net's skip connections and U-Net++'s dense pathways complement each other for fine-grained boundary detection.

DCAMF mask-level efficiency: The proposed DCAMF is lightweight, applied post-training, and avoids the computational cost of feature-level fusion, making it scalable and adaptable to new datasets.

Compared to previous aerial semantic segmentation works, the proposed DCAMF shows clear improvements: Hussein and Ali [20] achieved ~77% pixel accuracy (PA) with standard U-Net, Patil et al. [22] reported 82% PA using a hybrid CNN, and Maithil and Rehman [14] reached 86.1% PA with DensePlusU-Net. In contrast, the proposed DCAMF attained 88.08% PA and 0.671 mIoU on aerial data, confirming that fusion-based ensembles outperform single deep learning models for heterogeneous, multi-class aerial imagery.

Figure 17
Some examples of the segmentation results of the U-Net architecture on the second aerial dataset

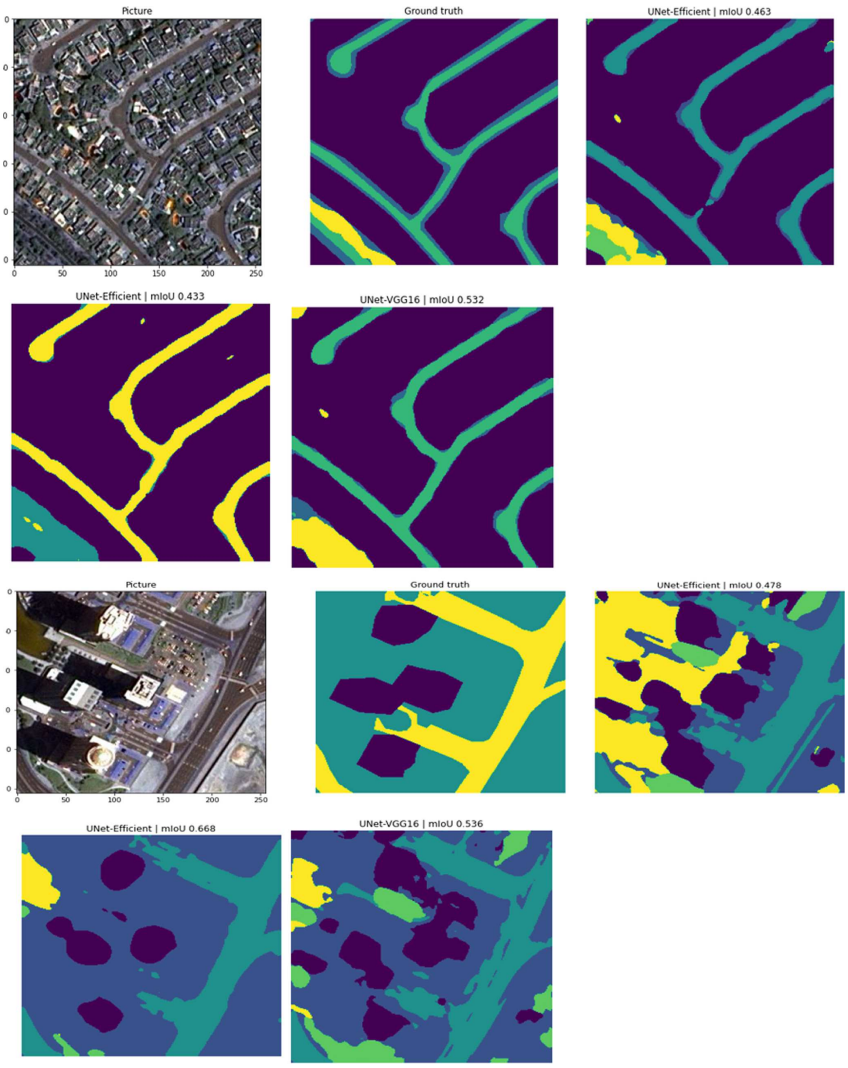


Table 6
Comparison between all U-Net models and the DCAMF models on the second aerial dataset

Model	Test accuracy (%)	Test mIoU
U-Net MobileNetV2	87.19	0.63
U-Net VGG16	85.7	0.62
U-Net EfficientNet	87.04	0.628
U-Net ResNet50	81.8	0.568
DCAMF (Top-3 U-Net Backbones)	88.03	0.6524
DCAMF (All U-Net Backbones)	88.08	0.6547

Figure 18
An example of the segmentation results of the best three individual models and DCAMF models

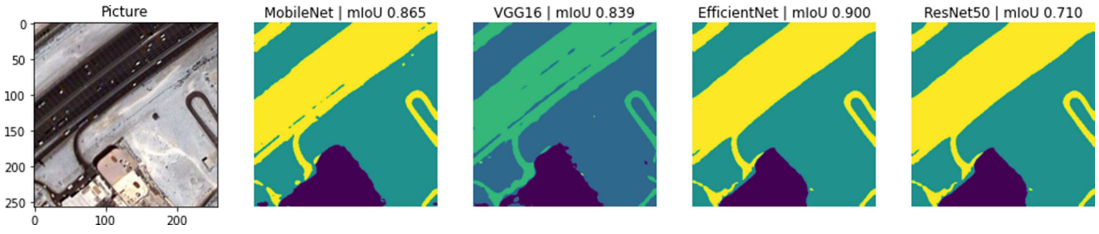


Table 7
Comparison between individual U-Net++ models and the DCAMF model

Model	Test accuracy (%)	Test mIoU
U-Net++ MobileNetV2	86.16	0.619
U-Net++ EfficientNet	88.05	0.642
DCAMF (All U-Net++ Backbones)	85.46	0.628

Figure 19
The loss, mIoU, and accuracy of training U-Net++ with MobileNetV2 on the second aerial dataset

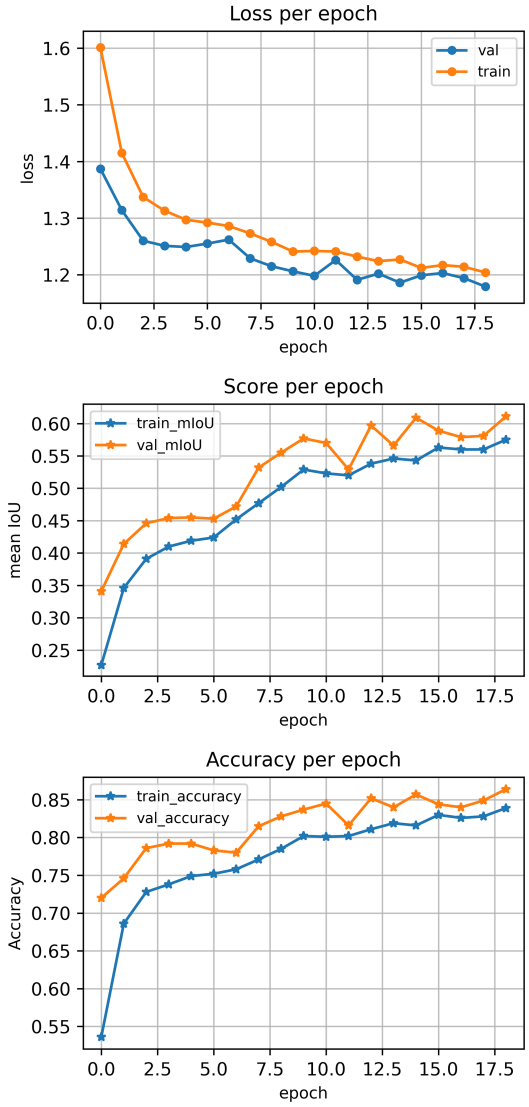
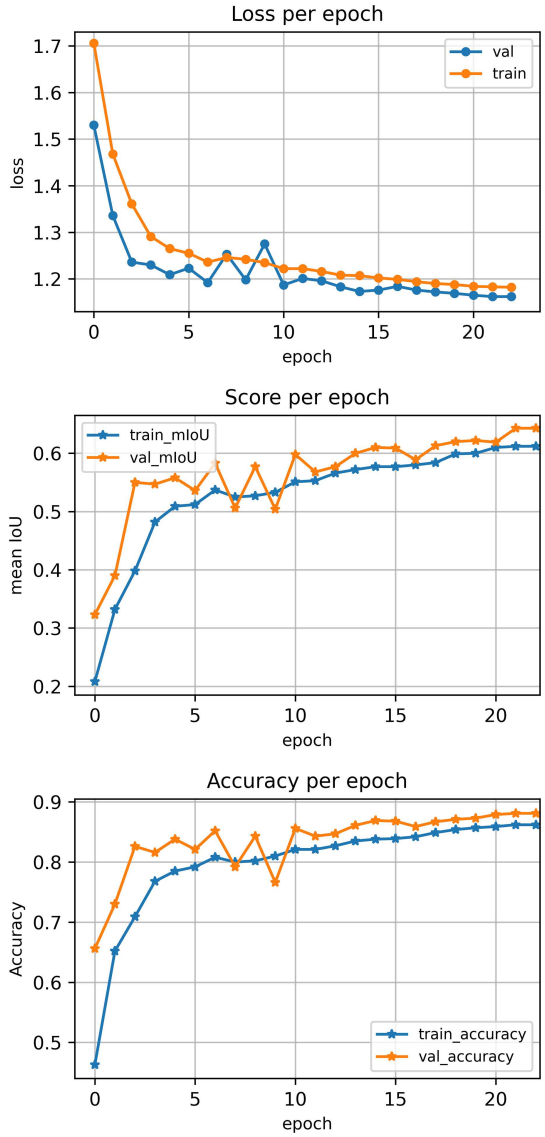


Figure 20
The loss, mIoU, and accuracy of training U-Net++ with EfficientNet on the second aerial dataset



4.7. Computational efficiency and resource requirements

Although the proposed framework involves multiple parallel models, the computational overhead remains manageable due to two factors: (1) lightweight backbones such as MobileNetV2 significantly reduce inference complexity, and (2) fusion is performed at the output mask level, avoiding intermediate feature recomputation. On the primary dataset, the average inference time per image is 0.42s for the full DCAMF ensemble compared to 0.21s for a single U-Net-MobileNetV2 model. GPU utilization increased by approximately 35%, confirming that the proposed framework, while costlier than standalone models, remains feasible for deployment in moderate-resource environments.

While the proposed method demonstrates performance improvements, it presents limitations. First, computational cost increases proportionally with the number of fused models, which may constrain deployment on edge devices. Second, generalization across domains remains dependent on dataset diversity; training from scratch may be required on unseen aerial environments. Third, DCAMF currently operates at the prediction level, and extending fusion to intermediate feature maps may further enhance performance but requires architectural redesign.

Figure 21
Some examples of segmentation results using the U-Net++ model on the drone dataset

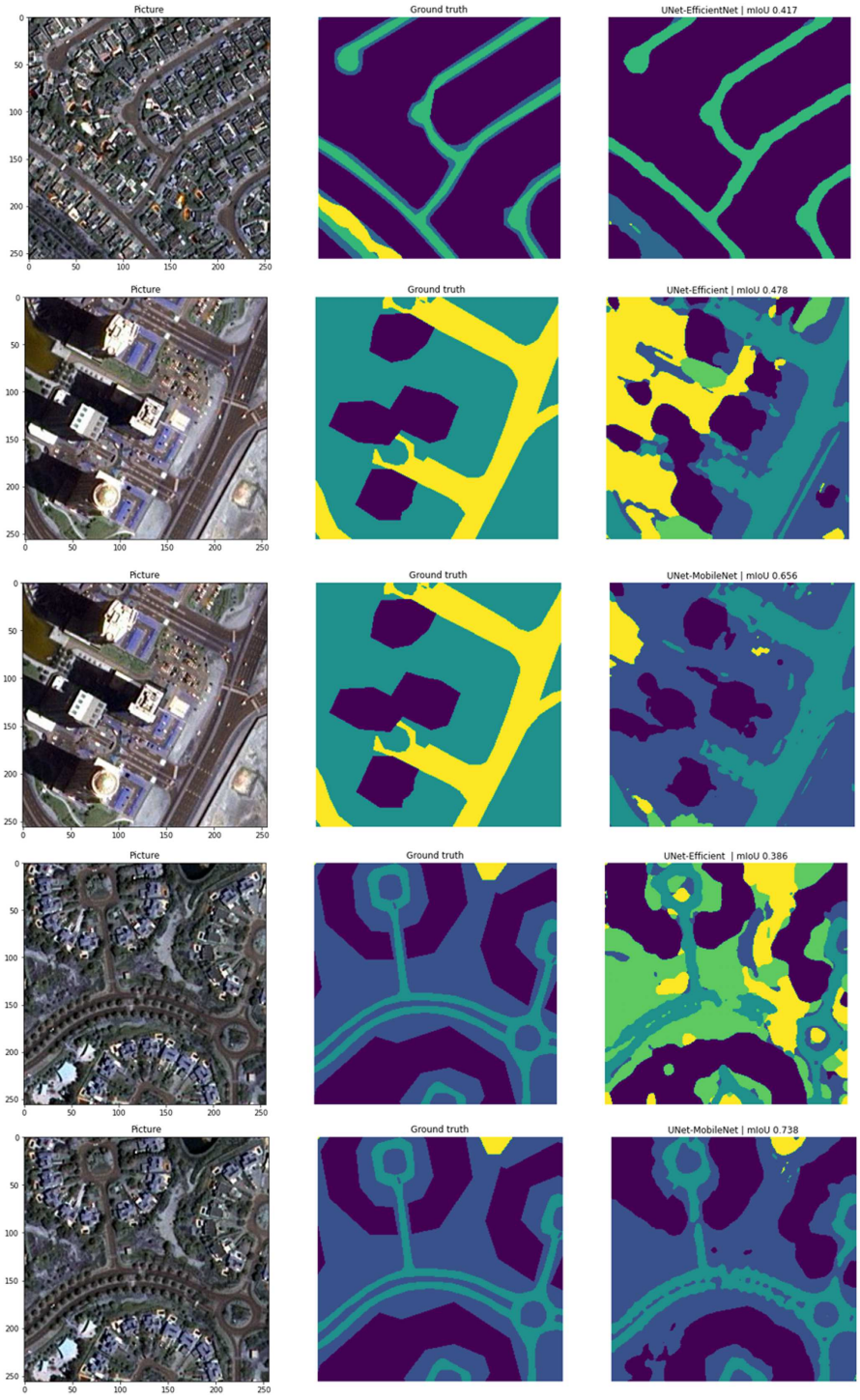


Table 8
Comparison of the DCAMF of U-Net, U-Net++, U-Net, and U-Net++ models on the aerial dataset

Model	Test accuracy (%)	Test mIoU
DCAMF (U-Net Backbones)	88.08	0.6547
DCAMF (U-Net++ Backbones)	85.46	0.628
DCAMF (U-Net + U-Net++)	88.06	0.671

Figure 22
Example of DCAMF of U-Net, U-Net++, U-Net, and U-Net++ models on the aerial dataset. (a) Example of DCAMF U-Net++ models. (b) Example of DCAMF U-Net and U-Net++ models

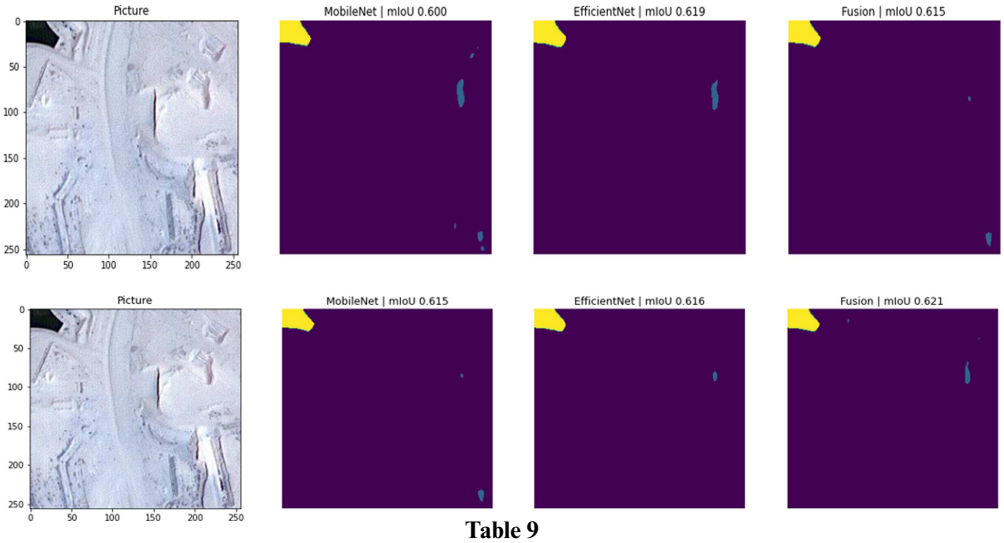


Table 9
A detailed comparison between the current study and previous works

Researcher and date	Methodology	Dataset	Results	Main limitations
[22]	Circle-U-Net	First dataset, drone dataset (400 images, 23 categories)	mIoU: 0.4647–0.5891	Low accuracy
[22]	U-Net	Second dataset, aerial dataset (72 images, 6 categories)	Pixel accuracy 82%	No other segmentation metrics rather than pixel accuracy
[14]	DensePlus U-Net	Second dataset, aerial dataset (72 images, 6 categories)	Pixel accuracy 86.11%	Additional dense connections require much more training time
[20]	U-Net	Second dataset, aerial dataset (72 images, 6 categories)	Pixel accuracy 76–78% mIoU: 0.52–0.55	Low accuracy
[23]	CNN U-Net	DJI PHANTOM four UAV drone (474 images)	The pixel accuracy was 91% and mIOU was 97%	Single category dataset
[26]	RFCN, YOLO, RetinaNet, U-Net, Conditional Random Fields	90197 (RGB images from Seagull dataset and 45200 images from Kaggle dataset	The IOU accuracy was 91% and 94%. ResNet-18 with U-Net.)	A two-step process, ResNet-18 with U-Net. Low detection accuracy
[16]	Mask R-CNN	INRIA-Austin, Vienna (72 images, 3528 sub-images)	The IOU accuracy is between 78.12% and 82.05%	High computational time
[27]	Custom Deep Neural Network (DNN), preprocessing	995 images of Qi Chen 2018 dataset	The pixel accuracy was 86%	Road semantic segmentation in satellite images (one category)
[28]	CNN U-Net	444 images (various categories)	The pixel accuracy was 85.8%	Low accuracy for injured person category
Proposed technique 2026	U-Net, U-Net++, DCAMF	First dataset, drone dataset (400 images, 23 categories) Second dataset, aerial dataset (72 images, 6 categories)	89.25% (first dataset), 88.08% (second dataset); mIoU: 0.5143 (first dataset), 0.671 (second dataset)	Improved performance using fusion techniques

5. Conclusion

This paper presented DCAMF, a robust segmentation framework that integrates U-Net and U-Net++ architectures through a mathematically grounded, uncertainty-aware fusion strategy. Unlike conventional ensemble approaches, DCAMF formulates mask fusion as a pixel-level decision process guided by entropy-based confidence estimation, enabling adaptive weighting of model predictions during inference.

Extensive experiments on aerial and drone image datasets with varying class complexities demonstrated that DCAMF consistently improves segmentation performance in terms of mIoU and overall accuracy, particularly in visually ambiguous regions and complex object boundaries. These results confirm that explicitly modeling prediction uncertainty plays a critical role in enhancing segmentation robustness.

Despite its effectiveness, DCAMF introduces additional computational overhead due to the use of multiple models during inference. Future work will focus on optimizing the fusion process and extending the proposed framework to transformer-based architectures and cross-domain aerial datasets.

Acknowledgement

The authors would like to thank Jadara University for supporting this work.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by any of the authors.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

Author Contribution Statement

Ahmad Qasim AlHamad: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization. **Ahmad Mousa Odat:** Validation. **Abdullah Ahmad Al Sokkar:** Supervision. **Mohammed Abdallah Otair:** Conceptualization, Methodology, Resources, Writing – review & editing. **Suhier Nadi Odah:** Formal analysis. **Omar Hussain Tarawneh:** Project administration.

References

[1] Khan, M. Z., Gajendran, M. K., Lee, Y., & Khan, M. A. (2021). Deep neural architectures for medical image semantic segmentation. *IEEE Access*, *9*, 83002–83024. https://doi.org/10.1109/ACCESS.2021.3086530

[2] Niu, R., Sun, X., Tian, Y., Diao, W., Feng, Y., & Fu, K. (2022). Improving semantic segmentation in aerial imagery via graph reasoning and disentangled learning. *IEEE Transactions on Geoscience and Remote Sensing*, *60*, 5611918. https://doi.org/10.1109/TGRS.2021.3121471

[3] Otair, M., Abualigah, L., Tawfiq, S., Alshinwan, M., Ezugwu, A. E., Zitar, R. A., & Sumari, P. (2024). Adapted arithmetic optimization algorithm for multi-level thresholding image segmentation: A case study of chest x-ray images. *Multimedia Tools and Applications*, *83*(14), 41051–41081. https://doi.org/10.1007/s11042-023-17221-9

[4] El-Shorbagy, M. A., Bouaouda, A., Nabwey, H. A., Abualigah, L., & Hashim, F. A. (2024). Advances in Henry gas solubility optimization: A physics-inspired metaheuristic algorithm with its variants and applications. *IEEE Access*, *12*, 26062–26095. https://doi.org/10.1109/ACCESS.2024.3365700

[5] Zhu, X. X., Tuia, D., Mou, L., Xia, G.-S., Zhang, L., Xu, F., & Fraundorfer, F. (2017). Deep learning in remote sensing: A comprehensive review and list of resources. *IEEE Geoscience and Remote Sensing Magazine*, *5*(4), 8–36. https://doi.org/10.1109/MGRS.2017.2762307

[6] Gupta, R., Jain, S., & Kumar, M. (2025). An enhanced algorithm for small object detection based on thermal imaging using YOLOv8-EPB. *Journal of Computer Science*, *21*(6), 1391–1403. https://doi.org/10.3844/jcssp.2025.1391.1403

[7] Zhang, C., Deng, X., & Ling, S. H. (2024). Next-gen medical imaging: U-Net evolution and the rise of transformers. *Sensors*, *24*(14), 4668. https://doi.org/10.3390/s24144668

[8] Li, Z., Zhang, H., Li, Z., & Ren, Z. (2022). Residual-attention UNet++: A nested residual-attention U-net for medical image segmentation. *Applied Sciences*, *12*(14), 7149. https://doi.org/10.3390/app12147149

[9] Isensee, F., Jaeger, P. F., Kohl, S., Petersen, J., & Maier-Hein, K. H. (2021). nnU-Net: Self-adapting framework for U-Net-based biomedical image segmentation. *Nature Methods*, *18*(2), 203–211. https://doi.org/10.1038/s41592-020-01008-z

[10] Hu, H., Cui, J., & Wang, L. (2021). Region-aware contrastive learning for semantic segmentation. In *2021 IEEE/CVF International Conference on Computer Vision*, 16271–16281. https://doi.org/10.1109/ICCV48922.2021.01598

[11] Wang, Z., Wang, B., Liu, Y., & Guo, J. (2023). Global feature attention network: Addressing the threat of adversarial attack for aerial image semantic segmentation. *Remote Sensing*, *15*(5), 1325. https://doi.org/10.3390/rs15051325

[12] Punn, N. S., & Agarwal, S. (2022). Modality specific U-Net variants for biomedical image segmentation: A survey. *Artificial Intelligence Review*, *55*(7), 5845–5889. https://doi.org/10.1007/s10462-022-10152-1

[13] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., …, & Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, *372*, n71. https://doi.org/10.1136/bmj.n71

[14] Maithil, K., & Rehman, T. B. (2022). Semantic segmentation of urban area satellite imagery using DensePlusU-Net. In *2022 IEEE International Conference on Current Development in Engineering and Technology*, 1–6. https://doi.org/10.1109/CCET56606.2022.10080484

[15] Tan, M., & Le, Q. (2021). EfficientNetV2: Smaller models and faster training. In *Proceedings of the 38th International Conference on Machine Learning*, 10096–10106.

[16] Huang, W., Liu, Z., Tang, H., & Ge, J. (2021). Sequentially delineation of rooftops with holes from VHR aerial images using a convolutional recurrent neural network. *Remote Sensing*, *13*(21), 4271. https://doi.org/10.3390/rs13214271

[17] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., & Luo, P. (2021). SegFormer: Simple and efficient design for semantic segmentation with transformers. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, 12077–12090.

[18] Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., . . . , & Zhou, Y. (2021). Transunet: Transformers make strong encoders for medical image segmentation. *MICCAI Workshops*, 61–71. https://doi.org/10.48550/arXiv.2102.04306

[19] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., . . . , & Guo, B. (2021). Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision*, 9992–10002. https://doi.org/10.1109/ICCV48922.2021.00986

[20] Hussein, S., & Ali, K. (2022). Semantic segmentation of aerial images using U-Net architecture. *Iraqi Journal for Electrical and Electronic Engineering*, *18*(1), 58–63. https://doi.org/10.37917/ijeee.18.1.7

[21] Sun, F., V, A. K., Yang, G., Zhang, A., & Zhang, Y. (2021). Circle-U-Net: An efficient architecture for semantic segmentation. *Algorithms*, *14*(6), 159. https://doi.org/10.3390/a14060159

[22] Patil, D., Patil, K., Nale, R., & Chaudhari, S. (2022). Semantic segmentation of satellite images using modified U-Net. In *2022 IEEE Region 10 Symposium* (pp. 1–6). https://doi.org/10.1109/TENSYMP54529.2022.9864504

[23] EL Amraoui, K., Ezzaki, A., Abanay, A., Lghoul, M., Hadri, M., Amari, A., & Masmoudi, L. (2022). An algorithm for crops segmentation in UAV images based on U-Net CNN model: Application to Sugarbeets plants. *ITM Web of Conferences*, *46*, 05002. https://doi.org/10.1051/itmconf/20224605002

[24] Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv Preprint:*. 1409.1556.

[25] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition*, 770–778. https://doi.org/10.1109/CVPR.2016.90

[26] Pires, C., Damas, B., & Bernardino, A. (2022). An efficient cascaded model for ship segmentation in aerial images. *IEEE Access*, *10*, 31942–31954. https://doi.org/10.1109/ACCESS.2022.3159667

[27] Trivedi, H., Sheth, D., Barot, R., & Shah, R. (2021). Road segmentation from satellite images using custom DNN. In *Proceedings of Second International Conference on Computing, Communications, and Cyber-Security: IC4S 2020*, 927–944. https://doi.org/10.1007/978-981-16-0733-2_66

[28] Martinez-Soltero, G., Alanis, A. Y., Arana-Daniel, N., & Lopez-Franco, C. (2020). Semantic segmentation for aerial mapping. *Mathematics*, *8*(9), 1456. https://doi.org/10.3390/math8091456

How to Cite: AlHamad, A. Q., Odat, A. M., Al Sokkar, A. A., Otair, M. A., Odah, S. N., & Tarawneh, O. H. (2026). A Dynamic Confidence-Aware Mask Fusion Framework for Aerial Image Segmentation Using U-Net and U-Net++. *Journal of Computational and Cognitive Engineering*. https://doi.org/10.47852/bonviewJCCE62028252