



Segmentasi Semantik Biner Banjir pada Citra UAV: Evaluasi *Multi-Seed* dan Studi Ablasi dengan *Attention U-Net*

Singgih Arya Mahendra*

Universitas Dian Nuswantoro,
Indonesia

Ricardus Anggi Pramunendar

Universitas Dian Nuswantoro,
Indonesia

***Corresponding author:**

Ricardus Anggi Pramunendar, Universitas
Dian Nuswantoro, Indonesia.

✉ ricardus.anggi@dsn.dinus.ac.id

Article Info:

Article history:

Received: July 14, 2026

Revised: August 20, 2026

Accepted: September 07, 2026

Available Online: October 15, 2026

Keywords:

Binary Semantic Segmentation;

Flood Detection;

Attention U-Net;

Focal Tversky Loss



This is an open access article under the
CC BY-SA license.

Copyright © 2026 by Author. Published
by Politeknik Siber Cerdika International

Kata Kunci:

Segmentasi Semantik Biner;

Deteksi Banjir;

Attention U-Net;

Focal Tversky Loss

Abstract

Background: Detecting flood areas from Unmanned Aerial Vehicle (UAV) imagery requires models that prioritize recall to minimize undetected flooded areas (false negatives). This is crucial because detection failures in disaster situations carry significantly more fatal risks than false alarms (false positives). However, models that focus on recall while also demonstrating proven stability across repeated evaluations remain rarely explored.

Objective: This study proposes a binary semantic segmentation approach that combines the Attention U-Net architecture, a pre-trained ResNet50 backbone, and the Focal Tversky Loss calculation ($\alpha = 0.3$; $\beta = 0.7$; $\gamma = 1.33$) to minimize detection failure rates.

Methods: The method was validated on the Flood Semantic Segmentation Dataset (FSSD) through repeated three-fold evaluation (multi-seed) and 2x2 ablation study statistical testing.

Results: The results demonstrate high and stable performance with an F1-score of 95.91%, IoU of 93.67%, and recall of 97.09% that consistently exceeds precision at 95.36%, with a very small standard deviation of 0.17% for both F1-score and IoU. The ablation study confirms that Focal Tversky Loss (FTL) is the primary driver of high recall.

Conclusion: The proposed method is proven stable and reproducible across multiple training runs. This work contributes a rigorously validated recall-oriented segmentation framework—integrating a 2x2 ablation study, multi-seed evaluation, and statistical testing—applicable to disaster warning systems that cannot tolerate false negatives.

Abstrak

Latar belakang: Deteksi area banjir dari citra Unmanned Aerial Vehicle (UAV) menghadapi tantangan class imbalance, batas genangan yang tidak beraturan, dan biaya kesalahan yang asimetris antara false negative dan false positive. Pada situasi bencana, kegagalan mendeteksi area tergenang (*false negative*) berisiko jauh lebih fatal dibandingkan alarm palsu (*false positive*). Namun, model yang secara sistematis memprioritaskan recall sekaligus terbukti stabil secara berulang masih jarang dieksplorasi.

Tujuan: Penelitian ini mengusulkan pendekatan segmentasi semantik biner yang memadukan arsitektur Attention U-Net, backbone ResNet50 pra-latih, dan perhitungan Focal Tversky Loss ($\alpha = 0.3$; $\beta = 0.7$; $\gamma = 1.33$) untuk menekan angka kegagalan deteksi.

Metode: Metode ini divalidasi pada Flood Semantic Segmentation Dataset (FSSD) melalui evaluasi berulang tiga kali (*multi-seed*) dan uji statistik studi ablasi 2x2.

Hasil: Hasilnya membuktikan kinerja yang tinggi dan stabil dengan F1-score 95,91%, IoU 93,67%, serta recall 97,09% yang konsisten

melampaui presisi sebesar 95,36%, dengan simpangan baku F1-score maupun IoU yang sangat kecil yaitu 0,17%. Studi ablasi menegaskan bahwa Focal Tversky Loss (FTL) merupakan penggerak utama tingginya recall.

Kesimpulan: Metode usulan terbukti stabil dan dapat direproduksi pada berbagai proses pelatihan. Penelitian ini berkontribusi pada pengembangan kerangka segmentasi berorientasi recall yang divalidasi secara ketat—mengintegrasikan studi ablasi 2×2, evaluasi multi-seed, dan pengujian statistik—sehingga cocok untuk sistem peringatan bencana yang tidak mentoleransi kegagalan deteksi (*false negative*).

To cite this article: Mahendra, S. A., & Pramunendar, R. A. (2026). Segmentasi semantik biner banjir pada citra UAV: Evaluasi multi-seed dan studi ablasi dengan Attention U-Net. *Journal of Business, Social and Technology*, 7(4), 2541–2557. <https://doi.org/10.59261/bustechno.v7i4.779>

PENDAHULUAN

Binary semantic segmentation merupakan salah satu permasalahan pokok dalam *computer vision*, khususnya pada analisis citra penginderaan jauh. Salah satu tantangannya adalah *class imbalance*, yaitu kondisi ketika jumlah piksel pada kelas target jauh lebih sedikit dibandingkan kelas lainnya. Akibatnya, model cenderung mempelajari kelas yang mendominasi citra sehingga kemampuan untuk mendeteksi piksel pada kelas target menjadi berkurang. Kondisi tersebut umumnya ditunjukkan dengan menurunnya nilai *recall* (Büttner et al., 2024; Salehi et al., 2017; Yeung et al., 2022).

Secara global, banjir merupakan salah satu bencana alam yang paling sering terjadi dan berdampak luas. Laporan Centre for Research on the Epidemiology of Disasters CRED (2023) mencatat bahwa banjir menyumbang sekitar 44% dari seluruh kejadian bencana alam di dunia dalam satu dekade terakhir, menimbulkan kerugian ekonomi mencapai miliaran dolar serta memengaruhi ratusan juta jiwa setiap tahunnya (CRED, 2023; Wallemacq & House, 2018). Dalam konteks penanganan bencana, penggunaan Unmanned Aerial Vehicle (UAV) telah berkembang pesat sebagai platform pencitraan yang efisien untuk pemetaan dan pengawasan wilayah terdampak secara real-time (Erdelj et al., 2017; Wibowo et al., 2025). Sejalan dengan itu, perkembangan deep learning pada citra penginderaan jauh telah membuka peluang baru untuk otomasi analisis citra bencana secara akurat dan skalabel (Li et al., 2022; X. X. Zhu et al., 2017).

Pada beberapa aplikasi, kesalahan berupa *false negative* memiliki konsekuensi yang lebih serius dibandingkan *false positive* (Salehi et al., 2017; R. Zhu et al., 2025). Artinya, piksel target yang tidak berhasil terdeteksi dapat menimbulkan dampak yang lebih besar daripada kesalahan mendeteksi objek yang sebenarnya tidak ada. Dalam kondisi seperti ini, *recall* menjadi metrik evaluasi yang lebih penting dibandingkan *precision* karena menunjukkan kemampuan model dalam menemukan sebanyak mungkin piksel target.

Salah satu contoh kasus yang memiliki karakteristik tersebut adalah deteksi banjir menggunakan citra *Unmanned Aerial Vehicle* (UAV). Pada citra banjir, area yang tergenang sering kali hanya menempati sebagian kecil dari keseluruhan piksel, memiliki bentuk dan batas yang tidak beraturan, serta sulit dibedakan dari objek di sekitarnya. Selain itu, kegagalan mendeteksi area banjir berpotensi menimbulkan dampak yang lebih besar dibandingkan kesalahan memberikan peringatan pada area yang sebenarnya tidak tergenang (*false alarm*) (Hernández et al., 2022; Simantiris & Panagiotakis, 2024; Sundaresan & Solomon, 2025). Oleh karena itu, kasus deteksi banjir dipilih sebagai dasar penyusunan penelitian ini karena mampu merepresentasikan permasalahan *class imbalance* sekaligus menekankan pentingnya peningkatan nilai *recall*.

Mencapai tingkat recall yang tinggi pada pendeteksian banjir tidaklah mudah. Tantangan utamanya adalah karakteristik objek air yang sering memiliki batas tidak tegas akibat variasi tekstur, bayangan, vegetasi, maupun keberadaan bangunan. Di samping itu, ketidakseimbangan (*class imbalance*) menambah sulitnya pencapaian recall walaupun derajatnya bergantung pada karakteristik dataset yang digunakan. Selain itu, fungsi *loss* konvensional seperti *cross-entropy* menganggap *false positive* dan *false negative* itu sama, sehingga tidak sesuai untuk aplikasi yang memerlukan tingkat recall tinggi.

Di sisi lain, perkembangan arsitektur deep learning telah mendorong kemajuan yang signifikan pada segmentasi semantik. U-Net memperkenalkan arsitektur *encoder-decoder* dengan *skip connection* yang mempertahankan informasi spasial selama proses rekonstruksi citra (Ronneberger et al., 2015; Zhang et al., 2023). Arsitektur ini kemudian berkembang melalui

pemanfaatan backbone residual seperti ResNet He et al. (2016) and Shafiq and Gu (2022), yang memungkinkan ekstraksi fitur lebih dalam sekaligus memanfaatkan *transfer learning* dari ImageNet He et al. (2016) and Shafiq and Gu (2022), sebagaimana telah diterapkan pada berbagai tugas klasifikasi citra (Abd Zaid et al., 2025; Thapa et al., 2023; Zhuang et al., 2021). *Attention U-Net* juga menambahkan *attention gate* pada setiap *skip connection* untuk menyaring fitur latar yang tidak relevan, sehingga model lebih fokus pada objek terbatas kompleks (Jiang et al., 2025; Schlemper et al., 2019). Meskipun demikian, peningkatan arsitektur saja belum cukup untuk memenuhi kebutuhan prioritas *recall* yang bersumber dari asimetri biaya *false negative*.

Upaya meningkatkan nilai *recall* dapat dilakukan melalui pemilihan fungsi *loss* yang sesuai. Salah satu fungsi *loss* yang banyak digunakan adalah *Tversky Loss*, karena mampu memberikan bobot yang berbeda terhadap *false positive* dan *false negative* sehingga model dapat diarahkan untuk lebih mengutamakan deteksi piksel target (Salehi et al., 2017; R. Zhu et al., 2025). Pengembangan dari metode tersebut adalah *Focal Tversky Loss* (FTL) yang menambahkan mekanisme *focal modulation*, sehingga proses pelatihan lebih berfokus pada piksel-piksel yang sulit diklasifikasikan (Abraham & Khan, 2019; Rema et al., 2026). Dengan pengaturan parameter yang tepat, FTL berpotensi meningkatkan kemampuan model dalam mendeteksi area banjir sehingga nilai *recall* menjadi lebih tinggi.

Meskipun demikian, terdapat sejumlah celah penelitian yang belum banyak dibahas pada studi-studi sebelumnya. Dari sisi kelebihan, studi sebelumnya berhasil menunjukkan efektivitas Focal Tversky Loss pada segmentasi medis dan keunggulan *attention gate* dalam menyaring fitur latar belakang (Abraham & Khan 2019; Jiang et al., 2025; Schlemper et al., 2019; Yeung et al., 2022). Namun, kelemahan yang teridentifikasi meliputi: (1) penerapan kombinasi FTL dan Attention U-Net pada segmentasi biner citra UAV masih sangat terbatas; (2) sebagian besar penelitian melaporkan hasil berdasarkan satu kali proses pelatihan (*single-run*) tanpa evaluasi reproduktibilitas; dan (3) validasi statistik antar-komponen melalui studi ablasi faktorial masih jarang dilakukan secara sistematis. Kekosongan inilah yang menjadi landasan utama penelitian ini.

Novelty utama penelitian ini terletak pada tiga aspek: (1) integrasi pertama Attention U-Net dengan backbone ResNet50 pra-latih dan Focal Tversky Loss pada segmentasi biner citra UAV untuk deteksi banjir secara berorientasi *recall*; (2) validasi stabilitas model melalui evaluasi multi-seed (tiga seed berbeda) disertai pelaporan statistik yang transparan; serta (3) studi ablasi faktorial 2×2 yang mengisolasi kontribusi masing-masing komponen secara statistik menggunakan uji Wilcoxon signed-rank berpasangan dengan koreksi perbandingan ganda.

Berdasarkan latar belakang tersebut, penelitian ini berfokus pada perancangan dan validasi pendekatan segmentasi semantik biner yang mengutamakan *recall* melalui studi kasus deteksi banjir berbasis citra UAV. Pendekatan yang diusulkan menggabungkan *Attention U-Net*, backbone ResNet50 pra-latih, dan Focal Tversky Loss sebagai fungsi *loss* yang dirancang untuk menekan *false negative*. Penelitian ini memberikan empat kontribusi utama, yaitu: (1) mengintegrasikan *attention gate*, *backbone residual pra-latih*, dan Focal Tversky Loss dalam satu kerangka segmentasi yang berorientasi pada *recall*; (2) melakukan studi ablasi faktorial 2×2 untuk mengisolasi kontribusi masing-masing komponen; (3) mengevaluasi kestabilan model melalui pengujian multi-seed yang disertai pelaporan statistik secara transparan; serta (4) memvalidasi pendekatan tersebut menggunakan *Flood Semantic Segmentation Dataset* (FSSD) (lihuayang111265, 2024) sebagai representasi kasus deteksi banjir yang kompleks.

Penelitian ini memiliki tiga implikasi utama yang relevan bagi bidang ilmu dan praktik: (1) bagi pengembangan sistem peringatan dini banjir, model yang secara konsisten memaksimalkan *recall* dapat mengurangi risiko area terdampak yang tidak terdeteksi; (2) bagi pemetaan pascabencana berbasis UAV, pendekatan ini menawarkan alat segmentasi otomatis yang andal dan dapat direproduksi; serta (3) bagi komunitas deep learning, penelitian ini menunjukkan bahwa evaluasi multi-seed dan studi ablasi faktorial merupakan protokol penting untuk memverifikasi stabilitas dan kontribusi komponen model secara ilmiah.

METODE

Penelitian ini merupakan penelitian kuantitatif eksperimental pada bidang computer vision. Pendekatan eksperimental digunakan karena penelitian dirancang untuk menguji hipotesis mengenai kontribusi komponen-komponen tertentu (attention gate dan fungsi loss) terhadap performa segmentasi melalui studi ablasi faktorial yang terkontrol, dan hasilnya dilaporkan secara numerik menggunakan metrik evaluasi standar.

Bagian ini menguraikan data, arsitektur, fungsi loss, prosedur pelatihan, dan skema evaluasi yang digunakan dalam penelitian. Seluruh nilai parameter yang dilaporkan mengikuti konfigurasi eksperimen sebenarnya agar dapat direproduksi.

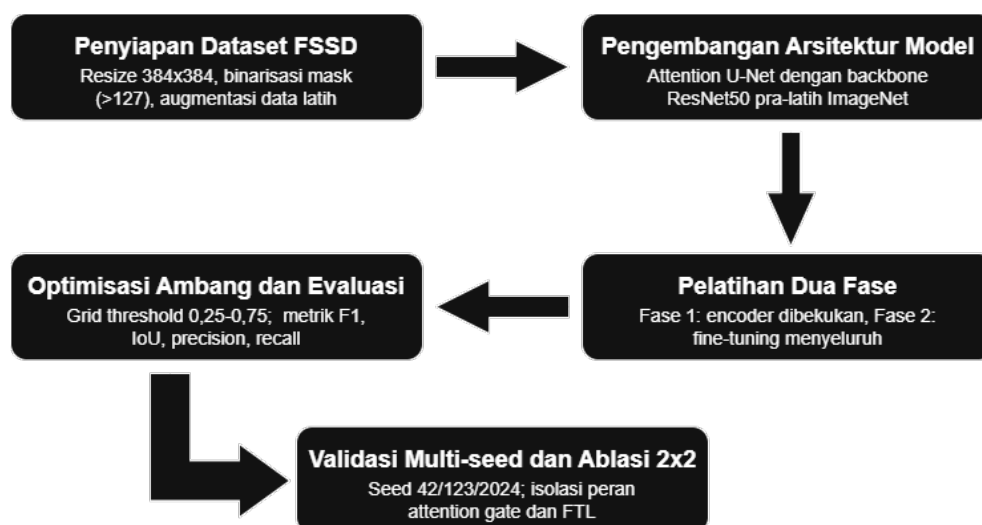
Kerangka Kerja Penelitian

Instrumen Penelitian: Seluruh eksperimen dilaksanakan menggunakan infrastruktur komputasi Google Colaboratory dengan GPU NVIDIA Tesla T4 (16 GB VRAM). Implementasi model menggunakan framework TensorFlow 2.x dengan library Keras API, didukung library NumPy, OpenCV, Albumentations (augmentasi), scikit-learn (evaluasi statistik), serta SciPy (uji Wilcoxon). Kode ditulis menggunakan Python 3.10. Tidak ada subjek manusia yang dilibatkan dalam penelitian ini; seluruh data yang digunakan merupakan dataset publik sehingga tidak memerlukan persetujuan etika penelitian berbasis subjek manusia.

Penelitian ini disusun melalui serangkaian tahapan yang saling berkaitan, dimulai dari penyiapan data hingga evaluasi akhir model. Tahap pertama adalah mempersiapkan *Flood Semantic Segmentation Dataset* (FSSD) (lihuayang111265, 2024) melalui proses pra-pemrosesan sehingga seluruh citra dan mask memiliki format yang seragam dan siap digunakan sebagai data masukan model. Dataset yang telah diproses kemudian digunakan untuk membangun model segmentasi berbasis *Attention U-Net* dengan ResNet50 sebagai backbone.

Setelah arsitektur selesai dirancang, proses pelatihan dilakukan dalam dua tahap. Tahap pertama memanfaatkan bobot pra-latih ResNet50 dengan membekukan seluruh lapisan encoder sehingga proses pembelajaran difokuskan pada decoder dan *attention gate*. Selanjutnya dilakukan *fine-tuning* dengan membuka seluruh lapisan jaringan agar parameter dapat disesuaikan secara menyeluruh terhadap karakteristik data banjir.

Model yang telah selesai dilatih tidak langsung digunakan untuk evaluasi. Terlebih dahulu dilakukan optimasi nilai ambang (*threshold*) guna memperoleh batas probabilitas yang menghasilkan kinerja segmentasi terbaik. Setelah *threshold* optimum diperoleh, performa model dievaluasi menggunakan beberapa metrik segmentasi. Untuk memastikan hasil yang diperoleh bersifat konsisten dan bukan disebabkan oleh variasi acak selama proses pelatihan, penelitian ini juga menerapkan evaluasi *multi-seed* serta studi ablasi menggunakan rancangan faktorial 2×2 . Keseluruhan tahapan penelitian disajikan pada Gambar 1.



Gambar 1. Kerangka Kerja Penelitian

Dataset dan Pra-pemrosesan

Penelitian ini menggunakan *Flood Semantic Segmentation Dataset* (FSSD) (lihuayang111265, 2024) sebagai sumber data utama. Dataset tersebut berisi citra udara beserta anotasi area banjir pada tingkat piksel sehingga sesuai untuk tugas semantic segmentation. Pembagian data mengikuti pembagian resmi yang disediakan oleh dataset, yaitu sebanyak 600 citra digunakan sebagai data pelatihan dan 63 citra sebagai data validasi.

Sebelum digunakan pada proses pelatihan, seluruh citra terlebih dahulu melalui tahap pra-pemrosesan agar memiliki ukuran dan format yang seragam. Citra masukan diubah menjadi resolusi 384×384 piksel menggunakan interpolasi bilinear. Sementara itu, mask anotasi diproses menggunakan interpolasi nearest-neighbor agar batas antar kelas tetap terjaga dan tidak mengalami pencampuran nilai piksel.

Mask hasil anotasi kemudian dikonversi menjadi citra biner menggunakan nilai ambang 127. Melalui proses ini setiap piksel hanya memiliki dua kemungkinan nilai, yaitu 0 untuk wilayah yang tidak tergenang banjir dan 1 untuk wilayah yang tergenang banjir. Selanjutnya seluruh citra dinormalisasi mengikuti skema ResNet50 pra-latih, yaitu dengan mengubah urutan kanal RGB menjadi BGR serta mengurangi nilai rata-rata (mean subtraction) berdasarkan statistik ImageNet.

Hasil analisis terhadap data menunjukkan bahwa rata-rata area banjir menempati sekitar 35,44% dari seluruh piksel pada data pelatihan. Persentase tersebut menunjukkan bahwa kelas banjir masih merupakan kelas minoritas, meskipun tingkat ketidakseimbangannya tidak terlalu ekstrem. Oleh karena itu, penelitian ini lebih menitikberatkan pada peningkatan recall, bukan semata-mata karena distribusi kelas, melainkan karena biaya kesalahan yang bersifat asimetris. Dalam konteks deteksi banjir, kegagalan mendeteksi area yang benar-benar tergenang (*false negative*) memiliki konsekuensi yang jauh lebih besar dibandingkan munculnya alarm palsu (*false positive*).

Untuk meningkatkan keragaman data sekaligus mengurangi risiko *overfitting*, augmentasi hanya diterapkan pada data pelatihan. Teknik augmentasi yang digunakan meliputi pencerminan horizontal dan vertikal dengan probabilitas masing-masing sebesar 0,5, rotasi acak dalam kelipatan 90°, random crop dengan fraksi 0,70–1,00 yang kemudian dikembalikan ke ukuran semula, serta augmentasi fotometrik berupa perubahan kecerahan maksimum sebesar 12 dan penyesuaian kontras pada rentang 0,9 hingga 1,1. Sebaliknya, data validasi tidak mengalami augmentasi sehingga hasil evaluasi benar-benar mencerminkan kemampuan model terhadap data yang belum pernah dilihat sebelumnya.

Arsitektur Model Usulan

Model yang diusulkan merupakan perpaduan antara *Attention U-Net* dan ResNet50 sebagai encoder. ResNet50 yang digunakan telah dilatih sebelumnya pada ImageNet He et al. (2016) and Shafiq & Gu (2022) sehingga mampu menyediakan representasi fitur yang lebih baik melalui mekanisme transfer learning. Lapisan klasifikasi pada bagian akhir ResNet50 dihilangkan karena penelitian ini hanya memanfaatkan kemampuan ekstraksi fiturnya.

Empat feature map perantara digunakan sebagai jalur *skip connection*, yaitu keluaran dari lapisan *conv1_relu*, *conv2_block3_out*, *conv3_block4_out*, dan *conv4_block6_out*, sedangkan keluaran *conv5_block3_out* berfungsi sebagai *bottleneck* jaringan.

Pada sisi decoder, proses rekonstruksi resolusi dilakukan secara bertahap melalui empat blok decoder. Setiap blok diawali dengan upsampling menggunakan interpolasi bilinear untuk menggandakan ukuran *feature map*. Hasil upsampling kemudian digabungkan dengan *feature map encoder* melalui mekanisme *attention gate* sebelum diproses oleh dua lapisan konvolusi 3×3 yang masing-masing diikuti *Batch Normalization* dan fungsi aktivasi ReLU.

Penggunaan *upsampling bilinear* dipilih sebagai pengganti konvolusi transpose untuk mengurangi kemungkinan munculnya artefak *checkerboard* Odena et al. (2016) and Lee & Kang (2023) pada hasil segmentasi. Setelah decoder terakhir, dilakukan satu kali proses upsampling tambahan sehingga resolusi keluaran kembali menjadi 384×384 piksel. Selanjutnya diterapkan konvolusi 3×3 yang diikuti *Batch Normalization* dan ReLU, kemudian diakhiri dengan konvolusi 1×1 beraktivasi sigmoid untuk menghasilkan peta probabilitas banjir satu kanal.

Attention gate yang digunakan mengikuti mekanisme *additive attention* (Jiang et al., 2025; Schlemper et al., 2019). Fitur dari jalur *skip connection* dan sinyal gating terlebih dahulu

diproyeksikan menggunakan konvolusi 1×1 , kemudian dijumlahkan dan dilewatkan ke fungsi aktivasi ReLU. Hasil tersebut diproyeksikan kembali menggunakan konvolusi 1×1 dengan aktivasi sigmoid untuk menghasilkan koefisien attention yang selanjutnya dikalikan dengan feature map *skip connection*. Melalui mekanisme ini jaringan dapat mengurangi respons pada wilayah latar belakang yang tidak relevan sekaligus memperkuat representasi fitur pada area banjir yang memiliki batas objek kompleks. Secara keseluruhan model yang diusulkan memiliki sekitar 33 juta parameter dengan ukuran masukan $384 \times 384 \times 3$ dan keluaran $384 \times 384 \times 1$.

Fungsi Loss dan Strategi Pelatihan

Pelatihan model menggunakan *Focal Tversky Loss* (FTL), yaitu fungsi *loss* yang dirancang untuk meningkatkan kemampuan model dalam mendeteksi kelas minoritas melalui penekanan terhadap kesalahan *false negative*. FTL dibangun berdasarkan indeks Tversky yang memberikan bobot berbeda pada false positive dan false negative sebagaimana ditunjukkan pada Persamaan (1).

$$(1) TI = \frac{TP + \varepsilon}{TP + \alpha \cdot FP + \beta \cdot FN + \varepsilon}$$

Pada penelitian ini digunakan $\alpha = 0,3$ dan $\beta = 0,7$ sehingga kesalahan *false negative* memperoleh penalti yang lebih besar dibandingkan *false positive*. Pemilihan bobot tersebut mengikuti konvensi notasi pada keluarga *focal loss* Yeung et al. (2022), sedangkan konsep pembobotan asimetris mengacu pada indeks Tversky (Salehi et al., 2017; R. Zhu et al., 2025).

Selanjutnya diterapkan mekanisme fokus melalui parameter γ sebagaimana ditunjukkan pada Persamaan (2) (Abraham & Khan, 2019; Rema et al., 2026).

$$(2) L_{FTL} = (1 - TI)^\gamma$$

Nilai γ ditetapkan sebesar 1,33 sehingga proses pelatihan lebih menitikberatkan pada piksel-piksel yang sulit diklasifikasikan. Kombinasi antara β yang lebih besar daripada α serta mekanisme fokus tersebut secara khusus diarahkan untuk meningkatkan nilai *recall*.

Pelatihan dilakukan menggunakan optimizer Adam dengan ukuran batch sebanyak delapan citra. Proses pelatihan dibagi menjadi dua fase. Pada fase pertama seluruh lapisan encoder ResNet50 dibekukan sehingga hanya decoder dan attention gate yang diperbarui. Tahap ini menggunakan learning rate sebesar 0,001 dengan maksimum 50 epoch. *Early stopping* diterapkan berdasarkan nilai IoU validasi dengan patience delapan *epoch*, sedangkan ReduceLROnPlateau menurunkan learning rate sebesar 50% apabila validation *loss* tidak mengalami perbaikan selama tiga *epoch* berturut-turut dengan batas minimum 10^{-6} .

Setelah fase pertama selesai, seluruh lapisan jaringan dibuka kembali sehingga seluruh parameter dapat disesuaikan melalui proses *fine-tuning*. Tahap kedua menggunakan learning rate yang jauh lebih kecil, yaitu 0,00001, agar bobot hasil *pre-training* tetap terjaga. Proses ini berlangsung maksimal 50 epoch dengan *early stopping* berpatience 12 epoch serta ReduceLROnPlateau berfaktor 0,5, patience lima epoch, dan batas minimum learning rate sebesar 10^{-7} .

Skema Evaluasi

Model menghasilkan keluaran berupa peta probabilitas sehingga diperlukan proses *thresholding* untuk memperoleh prediksi biner. Nilai *threshold* ditentukan melalui *grid search* pada rentang 0,25 hingga 0,75 dengan interval 0,05. *Threshold* yang menghasilkan nilai IoU agregat tertinggi pada data validasi kemudian digunakan sebagai *threshold* akhir. Kinerja model dievaluasi menggunakan empat metrik segmentasi, yaitu *Precision*, *Recall*, *F1-score*, dan *Intersection over Union* (IoU). Seluruh metrik dihitung berdasarkan jumlah *true positive* (TP), *false positive* (FP), dan *false negative* (FN) pada tingkat piksel sebagaimana ditunjukkan pada Persamaan (3) hingga Persamaan (6).

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

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

$$(5) F1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}$$

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

Untuk mengetahui kontribusi masing-masing komponen terhadap peningkatan performa model, penelitian ini melakukan studi ablasi menggunakan rancangan faktorial 2×2. Dua faktor yang dibandingkan adalah penggunaan *attention gate* (dengan atau tanpa *attention*) serta jenis fungsi *loss* (*Dice Loss* dan *Focal Tversky Loss*) (Mienye et al., 2025; Milletari et al., 2016). Kombinasi kedua faktor tersebut menghasilkan empat konfigurasi model, yaitu Varian A (tanpa *attention* + *Dice Loss*), Varian B (tanpa *attention* + *Focal Tversky Loss*), Varian C (*Attention* + *Dice Loss*), dan Varian D (*Attention* + *Focal Tversky Loss*). Varian D merupakan model usulan dalam penelitian ini.

Seluruh konfigurasi dilatih menggunakan prosedur yang sama sehingga perbedaan performa hanya dipengaruhi oleh faktor yang diuji. Selain itu, untuk memastikan bahwa perbedaan hasil benar-benar signifikan secara statistik, nilai *Precision*, *Recall*, *F1-score*, dan *IoU* juga dihitung pada tingkat per-citra terhadap 63 data validasi. Selanjutnya setiap pasangan model dibandingkan menggunakan uji *Wilcoxon signed-rank* berpasangan dengan koreksi perbandingan ganda sehingga dapat dibedakan apakah selisih performa yang diperoleh berasal dari efek nyata atau hanya merupakan variasi acak selama proses pelatihan.

HASIL DAN PEMBAHASAN

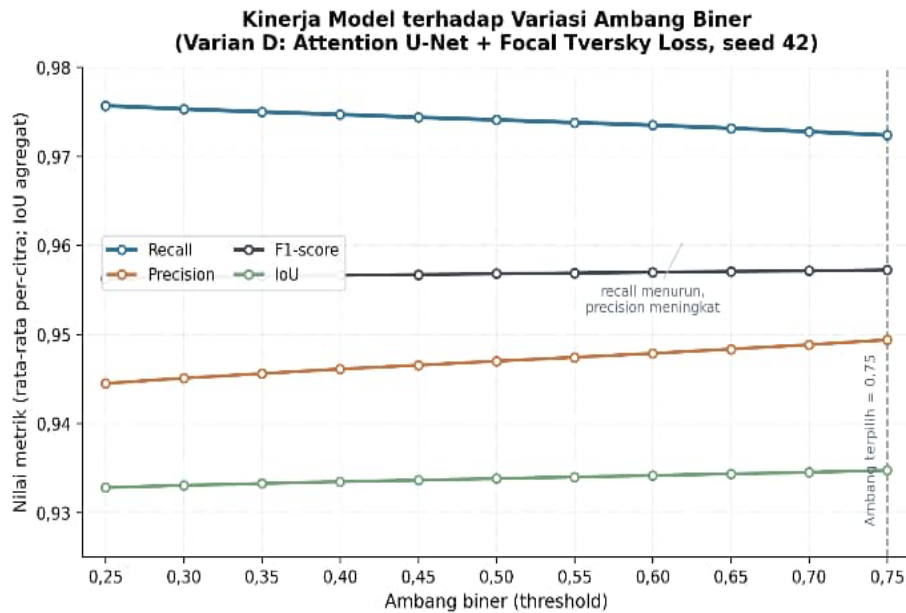
Bagian ini menyajikan hasil evaluasi model usulan secara bertahap, dimulai dari analisis pengaruh nilai ambang biner terhadap hasil segmentasi, dilanjutkan dengan evaluasi stabilitas model melalui pengujian *multi-seed*, studi ablasi untuk mengidentifikasi kontribusi setiap komponen penyusun model, serta analisis kualitatif terhadap hasil segmentasi. Seluruh temuan tersebut kemudian dibahas secara komprehensif untuk menjelaskan karakteristik, kemampuan, dan keterbatasan model dalam mendeteksi area banjir pada citra UAV.

Pengaruh Variasi Ambang Biner

Sebelum mengevaluasi kinerja akhir model, terlebih dahulu dianalisis pengaruh variasi nilai ambang biner (*threshold*) terhadap hasil segmentasi. Analisis ini bertujuan memastikan bahwa performa model tidak bergantung pada satu nilai ambang tertentu sehingga pemilihan *threshold* dilakukan berdasarkan pertimbangan yang objektif.

Seiring kenaikan nilai ambang dari 0,25 hingga 0,75, *precision* meningkat secara bertahap, sedangkan *recall* menurun dengan laju yang relatif landai. Di sisi lain, *F1-score* dan *Intersection over Union* (*IoU*) tetap berada pada kisaran yang hampir sama sepanjang rentang ambang yang diuji. Kestabilan kedua metrik agregat tersebut menunjukkan bahwa perubahan *threshold* tidak memberikan pengaruh yang berarti terhadap kualitas segmentasi secara keseluruhan.

Berdasarkan hasil tersebut, *threshold* sebesar 0,75 dipilih sebagai nilai operasional karena menghasilkan nilai *IoU* tertinggi pada data validasi tanpa menurunkan metrik lain secara signifikan. Menariknya, *recall* tetap berada di atas *precision* pada seluruh rentang *threshold*. Temuan ini menunjukkan bahwa karakteristik model yang berorientasi pada *recall* terutama berasal dari penggunaan *Focal Tversky Loss*, bukan semata-mata akibat pemilihan nilai *threshold*.



Gambar 2. Kinerja Varian D Terhadap Variasi Ambang Biner

Kinerja Model Usulan pada Evaluasi Multi-seed

Setelah threshold terbaik diperoleh, model usulan (Varian D) dievaluasi menggunakan tiga inisialisasi acak (*seed*) yang berbeda. Evaluasi ini dilakukan untuk memastikan bahwa performa model tidak bergantung pada satu proses pelatihan tertentu, sehingga hasil yang diperoleh dapat direproduksi secara konsisten.

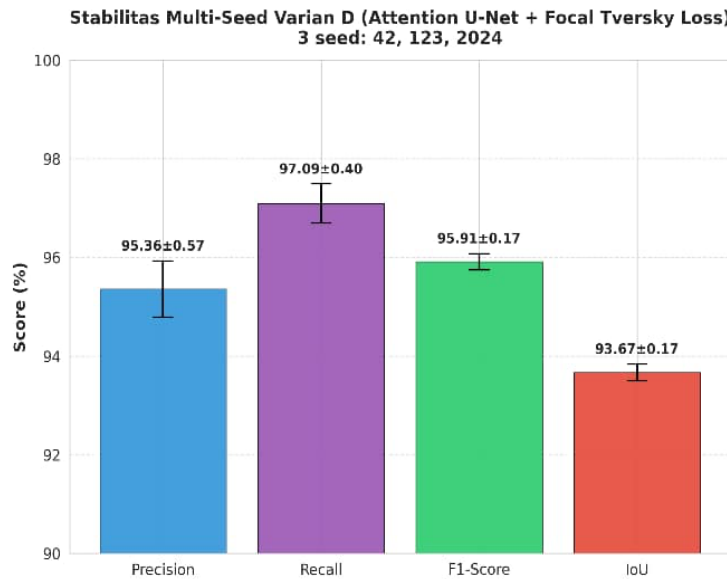
Tabel 1

Kinerja Varian D pada tiga seed beserta reratanya

Seed	Precision (%)	Recall (%)	F1 (%)	IoU (%)
42	94,94	97,24	95,72	93,47
123	95,13	97,40	95,97	93,74
2024	96,01	96,64	96,04	93,79
Rerata ± SB	95,36 ± 0,57	97,09 ± 0,40	95,91 ± 0,17	93,67 ± 0,17

Hasil pengujian pada ketiga seed disajikan pada Tabel 1. Secara rata-rata model mencapai *precision* sebesar 95,36%, *recall* 97,09%, *F1-score* 95,91%, dan *IoU* 93,67%. Pada seluruh seed, nilai *recall* selalu lebih tinggi dibandingkan *precision*. Pola tersebut menunjukkan bahwa model secara konsisten lebih memilih mendeteksi sebanyak mungkin piksel banjir meskipun harus menerima sejumlah kecil kesalahan positif (*false positive*). Karakteristik ini sesuai dengan tujuan penelitian, yaitu meminimalkan kemungkinan area banjir tidak terdeteksi. Selain memiliki performa yang tinggi, model juga menunjukkan kestabilan yang baik. Simpangan baku hanya mencapai 0,17% untuk *F1-score* maupun *IoU*, sedangkan interval kepercayaan 95% pada *F1-score* berada pada rentang 95,49% hingga 96,33%. Nilai simpangan baku yang kecil mengindikasikan bahwa variasi akibat inisialisasi acak sangat terbatas sehingga proses pelatihan dapat dianggap stabil.

Sebaran performa pada ketiga seed divisualisasikan pada Gambar 3, yang memperlihatkan konsistensi hasil antarproses pelatihan.



Gambar 3. Multi-seed Error Bar Varian D

Studi Ablasi 2×2

Setelah kestabilan model dikonfirmasi, dilakukan studi ablasi untuk mengidentifikasi kontribusi masing-masing komponen penyusun model. Empat kombinasi arsitektur dibangun melalui perpaduan penggunaan Attention Gate dan fungsi loss, sehingga pengaruh kedua komponen tersebut dapat diamati secara terpisah. Seluruh hasil evaluasi dirangkum pada Tabel 2.

Tabel 2

Hasil studi ablasi 2×2 pada set validasi

Metrik / Parameter	Varian A	Varian B	Varian C	Varian D (Usulan)
Attention	Tanpa	Tanpa	Dengan	Dengan
Loss Function	Dice	Focal Tversky	Dice	Focal Tversky
Precision (%)	96,18	95,09	96,02	95,30
Recall (%)	96,16	96,92	96,51	96,78
F1-Score (%)	95,85	95,65	95,85	95,69
IoU (%)	93,60	93,20	93,62	93,25
Ambang (Threshold)	0,60	0,75	0,30	0,75

Pengaruh Fungsi Loss

Penggantian *Dice Loss* menjadi *Focal Tversky Loss* menghasilkan perubahan perilaku model yang konsisten. Pada kedua pasangan varian, *recall* meningkat secara signifikan, sedangkan *precision* mengalami penurunan. Pada pasangan tanpa attention (Varian A ke Varian B), *recall* meningkat sebesar 0,76% ($p < 0,001$), sementara *precision* turun sebesar 1,09% ($p < 0,001$). Pola serupa juga muncul pada pasangan dengan attention (Varian C ke Varian D), yaitu *recall* meningkat 0,26% ($p = 0,011$), sedangkan *precision* turun 0,72% ($p < 0,001$). Sebaliknya, perubahan pada *F1-score* relatif kecil, yaitu kurang dari 0,20 poin persen, dan tidak selalu signifikan secara statistik. Temuan ini memperlihatkan bahwa *Focal Tversky Loss* memang bekerja sebagaimana tujuan perancangannya, yaitu meningkatkan sensitivitas model terhadap piksel banjir melalui peningkatan *recall*, bukan mengejar nilai F1-score tertinggi.

Pengaruh Attention Gate

Berbeda dengan fungsi loss, pengaruh *Attention Gate* relatif lebih kecil dan bergantung pada jenis fungsi loss yang digunakan. Pada pasangan berbasis *Dice Loss* (Varian A ke Varian C), penambahan attention meningkatkan *recall* sebesar 0,36% ($p = 0,008$), sementara *precision* dan *F1-score* tidak mengalami perubahan yang bermakna.

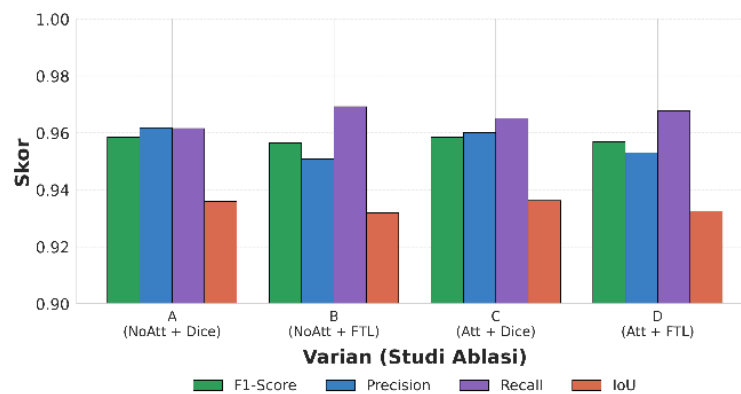
Sebaliknya, pada pasangan berbasis *Focal Tversky Loss* (Varian B ke Varian D), perhatian

utama *attention* bergeser menjadi menjaga *precision*, yang meningkat sebesar 0,21% ($p = 0,004$), sedangkan *recall* maupun *F1-score* tetap relatif sama. Dengan demikian, kontribusi utama Attention Gate bukan meningkatkan seluruh metrik secara seragam, melainkan membantu menyeimbangkan hubungan antara *precision* dan *recall*.

Pemilihan Model Usulan

Hasil studi ablasi menunjukkan bahwa setiap varian memiliki karakteristik yang berbeda sehingga tidak ada satu pun yang secara konsisten memberikan nilai terbaik pada seluruh metrik evaluasi. Varian A dan Varian C memperoleh *F1-score* serta *IoU* tertinggi, sedangkan Varian B menghasilkan *recall* tertinggi. Di sisi lain, Varian D mampu mempertahankan *recall* yang tinggi sekaligus mempertahankan penggunaan Attention Gate dengan penurunan *F1-score* yang sangat kecil dibandingkan varian berbasis *Dice*. Atas dasar tersebut, Varian D dipilih sebagai model usulan bukan karena menjadi yang terbaik pada seluruh metrik, melainkan karena paling sesuai dengan tujuan penelitian, yaitu menghasilkan model segmentasi yang berorientasi *recall* tanpa mengorbankan performa keseluruhan secara berarti.

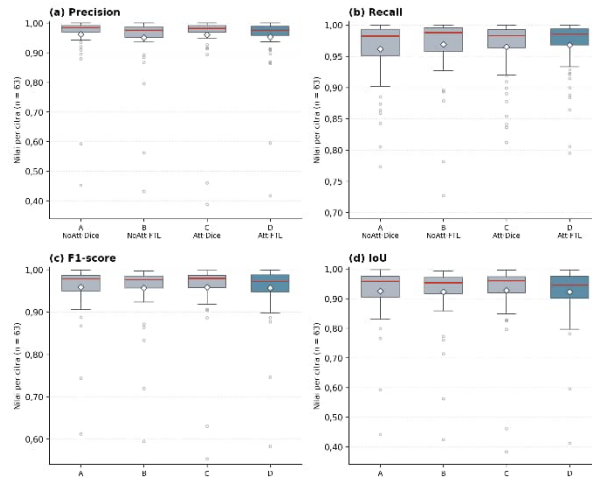
Uji Wilcoxon dilakukan terhadap hasil satu model terlatih untuk setiap varian sehingga signifikansi yang diperoleh menggambarkan variasi antar-citra validasi, bukan variasi akibat inisialisasi model. Selain itu, setelah koreksi perbandingan ganda diterapkan, efek Focal Tversky Loss terhadap *precision* tetap menjadi temuan yang paling kuat, sedangkan efek lain berada pada tingkat signifikansi yang lebih marginal.



Gambar 4. Perbandingan Ablasi

Sebaran metrik pada tingkat citra yang menjadi dasar pengujian statistik diperlihatkan pada Gambar 5. Sebagian besar citra memiliki nilai metrik yang tinggi, ditandai oleh median yang berada pada kisaran atas untuk seluruh varian. Meskipun demikian, masih terdapat beberapa penciran yang berasal dari citra dengan kondisi pemandangan yang kompleks.

Sebaran *recall* menunjukkan bahwa kedua varian berbasis *Focal Tversky Loss* memiliki median yang lebih tinggi dibandingkan varian berbasis *Dice Loss*. Pola tersebut konsisten dengan hasil uji *Wilcoxon* yang sebelumnya menunjukkan bahwa penggantian fungsi loss memberikan peningkatan *recall* secara signifikan.

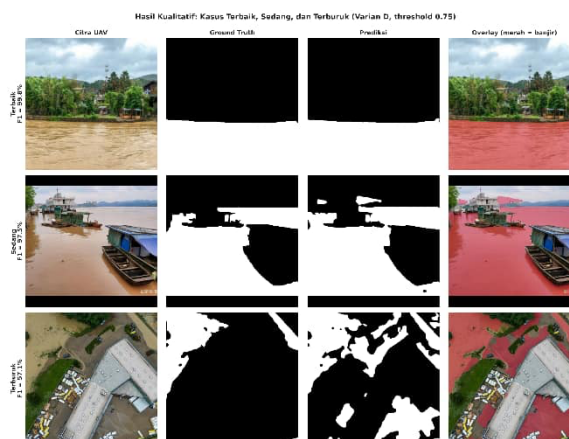


Gambar 5. Sebaran metrik per-citra antarvarian: (a) Precision, (b) Recall, (c) F1-score, (d) IoU

Analisis Kualitatif

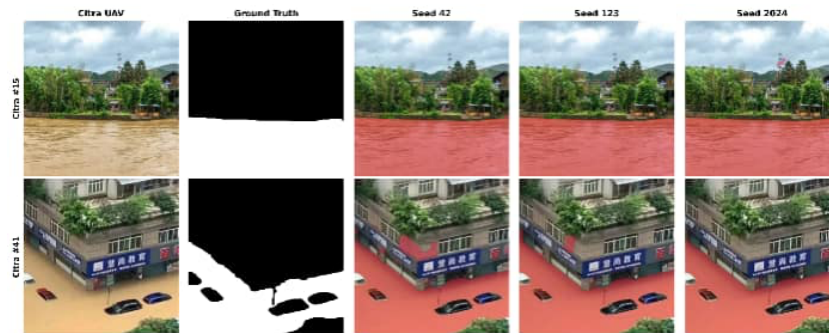
Meskipun metrik kuantitatif memberikan gambaran umum mengenai performa model, karakteristik kesalahan segmentasi lebih mudah dipahami melalui inspeksi visual terhadap hasil prediksi. Oleh karena itu, dilakukan analisis kualitatif pada beberapa citra yang mewakili kondisi terbaik, sedang, dan terburuk berdasarkan nilai *F1-score*. Pada kasus terbaik dengan *F1-score* sebesar 99,8%, area banjir yang luas dengan tekstur relatif homogen berhasil disegmentasi hampir sempurna. Batas genangan mengikuti kondisi sebenarnya, sedangkan vegetasi dan bangunan di sekitarnya berhasil dipisahkan dengan baik.

Pada kasus sedang dengan *F1-score* sebesar 97,5%, model tetap mampu mengenali area banjir meskipun terdapat objek lain seperti perahu dan kapal. Objek-objek tersebut sebagian besar berhasil dikecualikan sehingga kesalahan segmentasi hanya muncul pada sebagian kecil area tepi. Sebaliknya, pada kasus terburuk dengan *F1-score* sebesar 57,1%, model menghasilkan segmentasi yang berlebihan pada lingkungan dengan bangunan beratap luas dan area berlumpur di sekitarnya. Kesalahan yang muncul didominasi oleh *false positive*, sedangkan area banjir tetap berhasil terdeteksi. Karakteristik tersebut konsisten dengan orientasi model yang lebih mengutamakan *recall* dibandingkan *precision*.



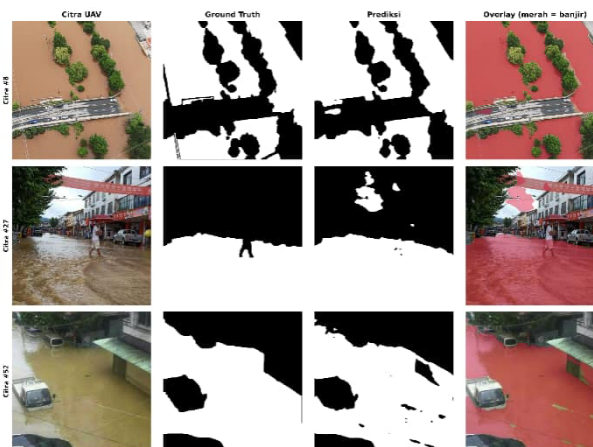
Gambar 6. Hasil segmentasi pada kasus terbaik, sedang, dan terburuk

Konsistensi keluaran model juga diamati melalui hasil segmentasi pada tiga seed yang berbeda. Gambar 7 memperlihatkan bahwa bentuk area banjir yang diprediksi hampir identik pada ketiga model. Hasil ini melengkapi evaluasi kuantitatif sebelumnya dan menunjukkan bahwa kestabilan model tidak hanya tercermin pada nilai metrik, tetapi juga pada konsistensi prediksi spasial.



Gambar 7. Konsistensi Hasil Segmentasi Antarseed

Untuk memberikan gambaran yang lebih luas mengenai perilaku model, Gambar 8 menyajikan beberapa contoh hasil segmentasi pada berbagai jenis pemandangan. Secara umum model mampu mengenali area banjir dengan baik, meskipun pada lingkungan yang kompleks masih ditemukan kecenderungan munculnya *false positive* sebagaimana telah dibahas sebelumnya.



Gambar 8. Contoh hasil segmentasi pada beragam skenario

Pembahasan

Hasil penelitian menunjukkan bahwa tujuan utama penelitian berhasil dicapai, yaitu mengembangkan model segmentasi area banjir yang berorientasi pada peningkatan *recall* tanpa mengorbankan kualitas segmentasi secara keseluruhan. Pada evaluasi *multi-seed*, model usulan memperoleh rata-rata *recall* sebesar 97,09%, lebih tinggi dibandingkan *precision* sebesar 95,36%, dengan *F1-score* 95,91% dan IoU 93,67% sebagaimana disajikan pada Tabel 1. Konsistensi nilai tersebut pada tiga inisialisasi acak yang berbeda menunjukkan bahwa karakteristik model tidak bergantung pada satu proses pelatihan tertentu. Simpangan baku yang kecil pada seluruh metrik semakin memperkuat bahwa performa model bersifat stabil dan dapat direproduksi.

Temuan tersebut memiliki makna yang penting dalam konteks pemetaan banjir berbasis citra UAV. Pada aplikasi kebencanaan, kesalahan berupa *false negative*, yaitu area banjir yang tidak berhasil terdeteksi, umumnya memiliki konsekuensi yang lebih besar dibandingkan *false positive*, karena dapat menyebabkan bagian wilayah terdampak tidak masuk dalam proses analisis maupun penanganan. Oleh karena itu, karakteristik model yang secara konsisten menghasilkan *recall* lebih tinggi dibandingkan *precision* menunjukkan bahwa strategi optimasi yang digunakan telah sesuai dengan tujuan penelitian.

Studi ablasi yang dirangkum pada Tabel 2 memperjelas sumber karakteristik tersebut. Hasil pengujian menunjukkan bahwa peningkatan *recall* terutama berasal dari penggunaan *Focal Tversky Loss*, sedangkan Attention Gate memberikan kontribusi yang lebih moderat. Penggantian *Dice Loss* menjadi *Focal Tversky Loss* secara konsisten meningkatkan *recall* pada kedua pasangan varian, meskipun diikuti sedikit penurunan *precision*. Sebaliknya, penambahan *Attention Gate* tidak meningkatkan seluruh metrik secara seragam, tetapi membantu menjaga keseimbangan antara *precision* dan *recall* sesuai fungsi *loss* yang digunakan.

Interpretasi tersebut diperkuat oleh hasil uji statistik yang telah dipaparkan pada bagian hasil. Efek *Focal Tversky Loss* terhadap peningkatan *recall* terbukti signifikan, sedangkan perubahan *F1-score* relatif kecil. Dengan demikian, fungsi loss ini tidak bekerja untuk memaksimalkan *F1-score*, melainkan memang menggeser perilaku model agar lebih sensitif terhadap piksel banjir. Di sisi lain, *Attention Gate* lebih berperan dalam memperhalus proses ekstraksi fitur sehingga kontribusinya lebih terlihat pada distribusi *precision* dan *recall* daripada peningkatan nilai metrik agregat.

Temuan kuantitatif tersebut selaras dengan analisis visual yang disajikan pada Gambar 6, Gambar 7, dan Gambar 8. Pada kasus terbaik, model mampu mengikuti batas genangan hampir sempurna dan berhasil memisahkan area banjir dari vegetasi maupun bangunan di sekitarnya. Pada kondisi sedang, model tetap mampu mempertahankan kualitas segmentasi meskipun terdapat objek lain, seperti perahu atau kapal, yang berada di atas permukaan air. Sebaliknya, pada pemandangan yang lebih kompleks, kesalahan model didominasi oleh false positive, terutama pada permukaan berlumpur dan atap bangunan yang memiliki karakteristik visual menyerupai genangan.

Pola kesalahan tersebut konsisten dengan hasil evaluasi kuantitatif yang menunjukkan dominasi *recall* terhadap *precision*. Dengan kata lain, ketika model mengalami kesulitan dalam membedakan objek, kecenderungan yang muncul bukanlah menghilangkan area banjir, melainkan memperluas prediksi area tergenang. Karakteristik tersebut sesuai dengan strategi optimasi menggunakan *Focal Tversky Loss* yang memang dirancang untuk mengurangi kemungkinan terjadinya *false negative*.

Selain akurasi, aspek implementasi juga menjadi pertimbangan penting dalam penelitian ini. Model usulan memiliki sekitar 33 juta parameter dan mampu memproses satu citra berukuran 384×384 piksel dalam waktu rata-rata sekitar 330 milidetik menggunakan GPU NVIDIA Tesla T4. Meskipun belum memenuhi kategori real-time, waktu inferensi tersebut masih memadai untuk mendukung alur kerja pemetaan banjir berbasis UAV yang pada praktiknya lebih banyak dilakukan secara offline setelah proses akuisisi citra selesai. Dengan demikian, model masih memiliki potensi yang baik untuk diterapkan pada proses pemetaan pascabencana maupun analisis wilayah terdampak.

Hasil penelitian ini juga sejalan dengan berbagai penelitian terdahulu. Berbagai studi menunjukkan bahwa *Focal Tversky Loss* efektif meningkatkan *recall* pada permasalahan segmentasi dengan ketidakseimbangan kelas, khususnya pada segmentasi medis (Abraham & Khan, 2019; Salehi et al., 2017; Yeung et al., 2022). Hasil studi ablasi pada penelitian ini memperlihatkan kecenderungan yang sama pada domain segmentasi banjir berbasis citra UAV, sehingga memperluas bukti bahwa karakteristik *Focal Tversky Loss* tetap konsisten ketika diterapkan pada objek dan lingkungan yang berbeda.

Secara praktis, temuan tersebut memiliki implikasi langsung bagi pengembangan sistem deteksi banjir berbasis UAV. Penggunaan *Focal Tversky Loss* dengan konfigurasi $\alpha = 0,3$ dan $\beta = 0,7$ terbukti secara konsisten menggeser perilaku model agar lebih sensitif terhadap piksel banjir, sehingga dapat diadopsi sebagai komponen standar pada sistem peringatan dini banjir yang tidak mentoleransi kegagalan deteksi. Bagi praktisi yang membutuhkan keseimbangan antara presisi dan sensitivitas, parameter $\gamma = 1,33$ memberikan mekanisme fokus tambahan yang dapat disesuaikan berdasarkan karakteristik dataset spesifik. Temuan ini memperluas bukti yang ada bahwa karakteristik *Focal Tversky Loss*—yang sebelumnya telah divalidasi pada domain segmentasi medis—bersifat general dan relevan untuk diterapkan pada domain penginderaan jauh khususnya analisis bencana banjir.

Demikian pula, kontribusi *Attention Gate* yang ditemukan pada penelitian ini selaras dengan berbagai penelitian segmentasi citra penginderaan jauh (Khan & Jung, 2024; Schlemper et al., 2019). Secara ilmiah, mekanisme additive attention yang digunakan dalam Attention U-Net bekerja dengan menekan respons pada wilayah latar belakang yang tidak relevan melalui koefisien attention yang dihitung secara adaptif. Hal ini menjelaskan mengapa pengaruh *Attention Gate* tidak bersifat seragam terhadap seluruh metrik, melainkan lebih terlihat dalam menyeimbangkan distribusi *precision* dan *recall* sesuai dengan fungsi loss yang digunakan. Dengan kata lain, *Attention Gate* tidak dirancang untuk memaksimalkan nilai metrik agregat, melainkan untuk memperhalus proses ekstraksi fitur pada objek terbatas kompleks—suatu

karakteristik yang sangat relevan untuk deteksi banjir dengan batas genangan yang tidak beraturan. Selain itu, pelaporan evaluasi melalui beberapa proses pelatihan (multi-seed) merupakan salah satu kontribusi metodologis penelitian ini sebagai respons terhadap meningkatnya perhatian komunitas ilmiah mengenai reproducibility dalam penelitian deep learning (Kattenborn et al., 2022; Maxwell et al., 2022).

Untuk memberikan gambaran mengenai posisi penelitian ini terhadap penelitian terdahulu, dilakukan perbandingan dengan metode yang menggunakan Flood Semantic Segmentation Dataset (FSSD) yang sama. Ringkasan perbandingan tersebut disajikan di Tabel 3.

Tabel 3

Perbandingan kinerja pada Flood Semantic Segmentation Dataset (FSSD)

Metode	Pendekatan	Dataset	F1 (%)
UFS-HT-REMSimantiris & Panagiotakis (2024)	Unsupervised	FSSD (663)	81,7
Attention U-Net + FTL (usulan)	Supervised	FSSD (600/63)	95,91

Sebagaimana terlihat pada Tabel 3, model usulan memperoleh *F1-score* sebesar 95,91%, lebih tinggi dibandingkan metode *UFS-HT-REM* yang melaporkan *F1-score* sebesar 81,7% pada dataset yang sama (Simantiris & Panagiotakis, 2024). Akan tetapi, perbedaan tersebut tidak dapat dipandang sebagai perbandingan langsung (*head-to-head*), karena kedua penelitian menggunakan paradigma pembelajaran yang berbeda. Metode *UFS-HT-REM* merupakan pendekatan *unsupervised* yang tidak memerlukan proses pelatihan, sedangkan penelitian ini menggunakan pendekatan *supervised* dengan pembagian data menjadi 600 citra pelatihan dan 63 citra validasi. Oleh karena itu, perbandingan tersebut dimaksudkan untuk menunjukkan posisi capaian penelitian dalam konteks dataset yang sama, bukan sebagai dasar untuk mengklaim keunggulan mutlak salah satu metode. Berdasarkan seluruh hasil yang diperoleh, kontribusi utama penelitian ini tidak terletak semata-mata pada peningkatan nilai metrik evaluasi, tetapi pada pengembangan pendekatan segmentasi yang secara sadar dioptimalkan untuk meningkatkan recall melalui kombinasi *Attention U-Net* dan *Focal Tversky Loss*, kemudian divalidasi menggunakan evaluasi statistik, studi ablasi, analisis kualitatif, serta pengujian *multi-seed*. Pendekatan evaluasi yang komprehensif tersebut memberikan keyakinan yang lebih tinggi bahwa performa model tidak hanya baik pada satu kali pelatihan, tetapi juga konsisten ketika proses pelatihan diulang.

Meskipun demikian, penelitian ini masih memiliki beberapa keterbatasan yang perlu dicatat. Pertama, optimasi threshold dan pelaporan metrik dilakukan menggunakan set validasi yang sama sehingga performa yang diperoleh masih merepresentasikan kinerja pada data validasi, bukan pada data uji independen. Walaupun demikian, hasil analisis pada Gambar 2 menunjukkan bahwa perubahan threshold hanya memberikan variasi yang sangat kecil pada *F1-score* maupun *IoU*, sehingga potensi bias optimistis diperkirakan relatif terbatas. Kedua, evaluasi masih dilakukan pada satu dataset (FSSD), sehingga kemampuan generalisasi model terhadap wilayah geografis, jenis sensor, maupun kondisi pencahayaan yang berbeda masih memerlukan pengujian lebih lanjut. Ketiga, studi ablasi hanya dijalankan menggunakan satu seed sehingga perbedaan kecil antarvarian perlu diinterpretasikan secara hati-hati. Keempat, perbandingan dengan penelitian lain masih bersifat indikatif karena adanya perbedaan paradigma pembelajaran dan protokol evaluasi. Kelima, model dengan sekitar 33 juta parameter dan kebutuhan GPU NVIDIA Tesla T4 memerlukan sumber daya komputasi yang cukup tinggi, sehingga penerapan langsung pada perangkat edge atau UAV dengan sumber daya terbatas (misalnya NVIDIA Jetson Nano atau perangkat sejenis) memerlukan optimasi tambahan seperti quantization, pruning, atau knowledge distillation sebelum dapat diimplementasikan secara real-time.

Berbagai keterbatasan tersebut sekaligus memberikan arah pengembangan penelitian selanjutnya. Penggunaan set uji independen, evaluasi lintas dataset, studi ablasi multi-seed, serta pengujian pada citra UAV dengan karakteristik yang lebih beragam diharapkan dapat memperkuat kemampuan generalisasi model sekaligus meningkatkan validitas kesimpulan yang diperoleh.

KESIMPULAN

Penelitian ini berhasil mengembangkan dan memvalidasi pendekatan segmentasi semantik biner berorientasi recall untuk deteksi area banjir pada citra UAV melalui integrasi Attention U-Net, backbone ResNet50 pra-latih, dan Focal Tversky Loss ($\alpha = 0,3$; $\beta = 0,7$; $\gamma = 1,33$). Evaluasi multi-seed menghasilkan precision 95,36%, recall 97,09%, F1-score 95,91%, dan IoU 93,67% dengan simpangan baku yang sangat kecil (0,17%), membuktikan bahwa dominasi recall terhadap precision bersifat konsisten dan dapat direproduksi. Studi ablasi faktorial 2×2 mengkonfirmasi bahwa Focal Tversky Loss merupakan penggerak utama peningkatan recall, sedangkan Attention Gate berperan menyeimbangkan hubungan precision-recall secara adaptif.

Kontribusi ilmiah utama penelitian ini tidak hanya terletak pada capaian metrik yang tinggi, tetapi pada penyajian protokol evaluasi yang komprehensif—mencakup studi ablasi faktorial, uji statistik Wilcoxon, evaluasi multi-seed, dan analisis kualitatif—yang meningkatkan kepercayaan terhadap reliabilitas dan reproduktibilitas hasil. Penelitian selanjutnya disarankan untuk menggunakan set uji independen, melakukan evaluasi lintas dataset, menerapkan studi ablasi multi-seed, serta mengeksplorasi strategi kompresi model (quantization, pruning) guna mendukung deployment pada perangkat edge dengan sumber daya komputasi terbatas.

UCAPAN TERIMA KASIH

Penulis mengucapkan terima kasih kepada Fakultas Ilmu Komputer, Universitas Dian Nuswantoro atas dukungan fasilitas dan sumber daya yang digunakan selama pelaksanaan penelitian ini. Penulis juga berterima kasih kepada penyedia layanan Google Colaboratory yang memungkinkan akses terhadap infrastruktur komputasi GPU untuk proses pelatihan dan evaluasi model.

PERNYATAAN KONTRIBUSI PENULIS

Seluruh penulis berkontribusi dalam perancangan dan pelaksanaan penelitian, persiapan data, pengolahan serta analisis hasil, dan penyusunan serta revisi naskah. Semua penulis telah membaca dan menyetujui versi akhir naskah.

DAFTAR PUSTAKA

- Abd Zaid, M. M., Mohammed, A. A., & Sumari, P. (2025). Remote sensing image classification using convolutional neural network (CNN) and transfer learning techniques. *Journal of Computer Science*, 21(3), 635–645. <https://doi.org/10.3844/jcssp.2025.635.645>
- Abraham, N., & Khan, N. M. (2019). A Novel Focal Tversky Loss Function With Improved Attention U-Net for Lesion Segmentation. *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*, 683–687. <https://doi.org/10.1109/ISBI.2019.8759329>
- Büttner, M., Schneider, L., Krasowski, A., Pitchika, V., Krois, J., Meyer-Lueckel, H., & Schwendicke, F. (2024). Conquering class imbalances in deep learning-based segmentation of dental radiographs with different loss functions. *Journal of Dentistry*, 148, 105063. <https://doi.org/10.1016/j.jdent.2024.105063>
- CRED, C. for R. on the E. of D. (2023). 2022 Disasters in numbers: Climate in action. Centre for Research on the Epidemiology of Disasters (CRED).
- Erdelj, M., Natalizio, E., Chowdhury, K. R., & Akyildiz, I. F. (2017). Help from the sky: Leveraging UAVs for disaster management. *IEEE Pervasive Computing*, 16(1), 24–32. <https://doi.org/10.1109/MPRV.2017.11>
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
- Hernández, D., Cecilia, J. M., Cano, J.-C., & Calafate, C. T. (2022). Flood detection using real-time image segmentation from unmanned aerial vehicles on edge-computing platform. *Remote Sensing*, 14(1), 223. <https://doi.org/10.3390/rs14010223>
- Jiang, C., Wang, Y., Yuan, Q., Qu, P., & Li, H. (2025). A 3D medical image segmentation network based on gated attention blocks and dual-scale cross-attention mechanism. *Scientific Reports*, 15(1). <https://doi.org/10.1038/s41598-025-90339-y>
- Kattenborn, T., Schiefer, F., Frey, J., Feilhauer, H., Mahecha, M. D., & Dormann, C. F. (2022). Spatially

- autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 5, 100018. <https://doi.org/10.1016/j.ophoto.2022.100018>
- Khan, B. A., & Jung, J.-W. (2024). Semantic segmentation of aerial imagery using U-Net with self-attention and separable convolutions. *Applied Sciences*, 14(9), 3712. <https://doi.org/10.3390/app14093712>
- Lee, H., & Kang, M. G. (2023). Infrared Image Deconvolution Considering Fixed Pattern Noise. *Sensors*, 23(6). <https://doi.org/10.3390/s23063033>
- Li, J., Hong, D., Gao, L., Yao, J., Zheng, K., Zhang, B., & Chanussot, J. (2022). Deep learning in multimodal remote sensing data fusion: A comprehensive review. *International Journal of Applied Earth Observation and Geoinformation*, 112. <https://doi.org/10.1016/j.jag.2022.102926>
- Maxwell, A. E., Bester, M. S., & Ramezan, C. A. (2022). Enhancing reproducibility and replicability in remote sensing deep learning research and practice. *Remote Sensing*, 14(22), 5760. <https://doi.org/10.3390/rs14225760>
- Mienye, I. D., Swart, T. G., Obaido, G., Jordan, M., & Ilono, P. (2025). Deep Convolutional Neural Networks in Medical Image Analysis: A Review. In *Information (Switzerland)* (Vol. 16, Number 3). <https://doi.org/10.3390/info16030195>
- Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. *2016 Fourth International Conference on 3D Vision (3DV)*, 565–571. <https://doi.org/10.1109/3DV.2016.79>
- Odena, A., Dumoulin, V., & Olah, C. (2016). Deconvolution and checkerboard artifacts. *Distill*, 1(10). <https://doi.org/10.23915/distill.00003>
- Rema, P., Moorthy, S., Bright, B., & BR, M. (2026). An Attention-Guided Residual 3D U-Net with Focal Tversky-Dice Loss for Multi-Modal Pancreatic Tumor Segmentation Using Synthetic Volumetric Imaging. *Cancer Treatment and Research Communications*, 101146.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015* (pp. 234–241). Springer. https://doi.org/10.1007/978-3-319-24574-4_28
- Salehi, S. S. M., Erdogmus, D., & Gholipour, A. (2017). Tversky loss function for image segmentation using 3D fully convolutional deep networks. In *Machine Learning in Medical Imaging (MLMI 2017)* (Vol. 10541, pp. 379–387). Springer. https://doi.org/10.1007/978-3-319-67389-9_44
- Schlemper, J., Oktay, O., Schaap, M., Heinrich, M., Kainz, B., Glocker, B., & Rueckert, D. (2019). Attention gated networks: Learning to leverage salient regions in medical images. *Medical Image Analysis*, 53, 197–207. <https://doi.org/10.1016/j.media.2019.01.012>
- Shafiq, M., & Gu, Z. (2022). Deep Residual Learning for Image Recognition: A Survey. In *Applied Sciences (Switzerland)* (Vol. 12, Number 18). <https://doi.org/10.3390/app12188972>
- Simantiris, G., & Panagiotakis, C. (2024). Unsupervised color-based flood segmentation in UAV imagery. *Remote Sensing*, 16(12), 2126. <https://doi.org/10.3390/rs16122126>
- Sundaresan, A. A., & Solomon, A. A. (2025). Post-disaster flooded region segmentation using DeepLabv3+ and unmanned aerial system imagery. *Natural Hazards Research*, 5(2), 363–371. <https://doi.org/10.1016/j.nhres.2024.12.003>
- Thapa, A., Horanont, T., Neupane, B., & Aryal, J. (2023). Deep learning for remote sensing image scene classification: A review and meta-analysis. *Remote Sensing*, 15(19), 4804. <https://doi.org/10.3390/rs15194804>
- Wibowo, A., Amri, I., Surahmat, A., & Rusdah, R. (2025). Leveraging artificial intelligence in disaster management: A comprehensive bibliometric review. *Jamba: Journal of Disaster Risk Studies*, 17(1). <https://doi.org/10.4102/jamba.v17i1.1776>
- Yeung, M., Sala, E., Schönlieb, C.-B., & Rundo, L. (2022). Unified focal loss: Generalising Dice and cross entropy-based losses to handle class imbalanced medical image segmentation. *Computerized Medical Imaging and Graphics*, 95, 102026. <https://doi.org/10.1016/j.compmedimag.2021.102026>
- Zhang, J., Zhang, Y., Jin, Y., Xu, J., & Xu, X. (2023). MDU-Net: multi-scale densely connected U-Net for biomedical image segmentation. *Health Information Science and Systems*, 11(1).

- <https://doi.org/10.1007/s13755-022-00204-9>
- Zhuang, F., Qi, Z., Duan, K., Xi, D., Zhu, Y., Zhu, H., Xiong, H., & He, Q. (2021). A comprehensive survey on transfer learning. *Proceedings of the IEEE*, 109(1), 43–76. <https://doi.org/10.1109/JPROC.2020.3004555>
- Zhu, R., Oda, M., Hayashi, Y., Kitasaka, T., Misawa, K., Fujiwara, M., & Mori, K. (2025). Skeleton-guided 3D convolutional neural network for tubular structure segmentation. *International Journal of Computer Assisted Radiology and Surgery*, 20(1). <https://doi.org/10.1007/s11548-024-03215-x>
- 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>
- Wallemacq, P., & House, R. (2018). Economic losses, poverty & disasters: 1998-2017. Centre for Research on the Epidemiology of Disasters (CRED); United Nations Office for Disaster Risk Reduction (UNDRR). <https://www.undrr.org/publication/economic-losses-poverty-disasters-1998-2017>