

Article

Assessing the Domain Gap Between Public Crack Datasets and UAV Imagery for Deep Learning-Based Crack Segmentation

Seungchan Lim ¹, Donggyu Kim ¹, Hyojin Kim ¹, Jaehoon Jeong ¹, Chuluong Choi ¹  and Hoyong Ahn ^{2,*}

¹ Major of Spatial Information Engineering, Division of Earth Environmental System Science, Pukyong National University, Busan 48513, Republic of Korea; sc6888@pukyong.ac.kr (S.L.); kevin8220@pukyong.ac.kr (D.K.); hyojini101@pukyong.ac.kr (H.K.); jjang403@gmail.com (J.J.); cuchoi@pknu.ac.kr (C.C.)

² Department of Agricultural Biology, Climate Change and Agro-Ecology Division, National Institute of Agricultural Sciences, Rural Development Administration, Jeonju 55365, Republic of Korea

* Correspondence: hyahn85@korea.kr

Abstract

Deep learning-based road crack segmentation can achieve high benchmark performance on public datasets, but its reliability in UAV-acquired field imagery remains uncertain because real road scenes contain various crack-like non-crack features. This study assessed the field transferability of a public-dataset-trained crack segmentation model using UAV-acquired road images. Nine semantic segmentation model configurations were trained and compared using integrated public crack datasets, and U-Net++ with a ResNet50 backbone was selected for UAV field application. After confidence-based data refinement and training data augmentation, the final model achieved precision, recall, Dice, and IoU values of 0.802, 0.787, 0.794, and 0.659, respectively, on the public validation dataset. Field evaluation was conducted on 148 UAV images and manually corrected UAV ground-truth patches generated from non-overlapping 8×8 image grids. The total UAV field evaluation produced precision, recall, Dice, and IoU values of 0.856, 0.806, 0.830, and 0.710, respectively. Major errors were associated with crack-like non-crack factors, including lane markings, manholes, repaired boundaries, shadows, surface contamination, and reflections. Threshold sensitivity analysis indicated that approximately 0.4–0.5 was a practical operating range under the tested UAV conditions. These results show that idealized public-dataset evaluation can overestimate real-world UAV performance and that field transferability evaluation is essential for reliable UAV-based pavement inspection.

Keywords: UAV imagery; road crack segmentation; deep learning; field transferability; domain gap; pavement inspection



Academic Editor: Andrea Prati

Received: 1 June 2026

Revised: 26 June 2026

Accepted: 3 July 2026

Published: 11 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and

conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Road pavement is continuously exposed to vehicle loading, weathering, material aging, and repeated thermal expansion and contraction, which gradually induce surface distress such as cracks, raveling, settlement, and potholes. Among these distress types, pavement cracks are one of the most common early-stage indicators of pavement deterioration. If cracks are not detected and repaired in a timely manner, they can propagate and provide pathways for water and debris infiltration, which may weaken pavement layers and accelerate premature deterioration [1–3]. Therefore, continuous pavement monitoring and early crack management are important for supporting maintenance planning and preventing minor surface defects from progressing into more serious pavement distress [4]. In this context, the objective and efficient identification of crack locations, continuity,

and spatial distributions can provide essential information for pavement maintenance decision-making.

Conventional pavement condition surveys are commonly conducted using vehicle-based inspections, manual visual inspections, and field measurements of crack width and depth [5,6]. Although these approaches allow inspectors to directly observe pavement conditions, they require considerable time and labor when applied repeatedly over large road networks. In addition, inspection results may vary depending on the experience and judgment of individual surveyors, making it difficult to record crack locations and shapes in a consistent and reproducible manner. To overcome these limitations, many studies have explored automated road crack detection using image data and computer vision techniques. Earlier image-processing-based approaches, such as thresholding, edge detection, and morphological processing, can extract crack patterns from pavement images, but their performance is often sensitive to illumination changes, shadows, noise, and complex pavement textures.

Deep learning-based semantic segmentation models are useful for pavement crack detection because they can identify crack regions at the pixel level and describe the geometry, continuity, and spatial distribution of cracks [7]. However, benchmark performance on public crack datasets does not necessarily guarantee comparable performance in real-world field imagery. Many public road crack datasets contain cropped or crack-centered road-surface images in which crack regions are relatively clear, whereas crack-like non-crack factors are limited. In contrast, Unmanned Aerial Vehicle (UAV)-acquired road images contain broader road-surface contexts, illumination variations, shadows, lane markings, repaired pavement boundaries, manholes, surface contamination, and other non-crack linear objects. These field-specific factors can produce visual patterns similar to cracks and may cause false-positive (FP) and false-negative (FN) detections [8–10]. This difference creates a domain gap between public crack datasets and UAV-acquired field imagery, where models trained under idealized or crack-centered conditions may not maintain the same level of performance in practical road inspection scenarios.

Motivated by this domain gap, this study aims to evaluate the field transferability of a crack segmentation model trained using public crack datasets when applied to real-world UAV imagery. Rather than simply reporting benchmark accuracy under idealized dataset conditions, this study investigates whether a model trained on public crack datasets can maintain practical detection performance in UAV-acquired road-surface images.

To achieve this objective, nine candidate semantic segmentation model configurations were trained and compared using integrated public crack datasets. The effect of training data augmentation was evaluated by comparing models trained with and without augmentation under the same validation conditions. A separate confidence-based filtering experiment was then conducted on the selected model to evaluate the effect of reliable sample selection. The trained model was finally applied to UAV images acquired from actual campus roads and parking areas for field transferability evaluation. Supplementary quantitative evaluation was conducted using manually corrected UAV field ground truth (GT) patch subsets, and major error factors were examined under real road-surface conditions. Threshold sensitivity was also analyzed because threshold selection can affect crack thickness, connectivity, and noise detection [11,12].

The main contributions of this study are as follows:

- (1) the domain gap between public crack datasets and UAV-acquired road imagery was explicitly evaluated for road crack segmentation;
- (2) the effects of training data augmentation and confidence-based filtering were separately assessed under controlled public dataset validation conditions; and

- (3) UAV field performance and major FP/FN factors caused by crack-like non-crack road-surface features were analyzed using UAV images and manually corrected field GT patches.

2. Materials and Methods

2.1. Overall Study Workflow

The overall workflow of this study is shown in Figure 1. The workflow consists of four main stages: public crack dataset preparation, model comparison with and without data augmentation, confidence-based filtering of the selected model, and UAV field application for transferability evaluation.

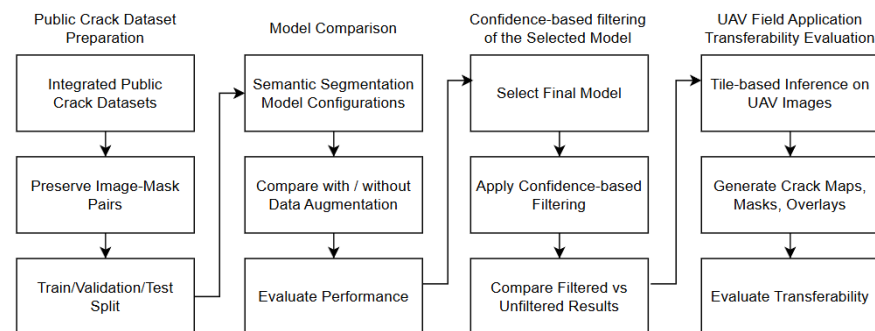


Figure 1. Overall workflow of model development, UAV field inference, and field transferability evaluation.

First, integrated public crack datasets were prepared for supervised model training and evaluation. Image–mask pairs were preserved during dataset construction, and the dataset was divided into training and validation subsets. UAV-acquired images were not used for model training but only for field application and transferability evaluation.

Second, nine candidate semantic segmentation model configurations were trained and compared under the same public dataset conditions. To evaluate the effect of training data augmentation, each model was trained with and without data augmentation. The results were used to assess model performance and the contribution of data augmentation.

Third, the final model was selected based on model comparison results, validation performance, training stability, and suitability for thin and discontinuous crack segmentation. For the selected model, confidence-based filtering was conducted to evaluate the effect of retaining high-confidence image–mask pairs. This effect was assessed by comparing models trained with and without filtering under the same validation conditions.

Finally, the selected model was applied to UAV-acquired road-surface images using patch-level processing based on a non-overlapping 8×8 grid to generate crack probability maps, predicted binary masks, and overlay images. Field transferability was evaluated using manually corrected UAV field ground-truth (GT) patches and expert visual assessment, including threshold sensitivity and major FP/FN error factor analysis under real-world UAV conditions.

2.2. Study Materials and UAV Image Acquisition

The research materials consisted of public crack segmentation datasets and UAV-acquired road surface images. The public datasets were used for model training, candidate model comparison, and benchmark validation. The UAV-acquired images were used to evaluate field transferability under real-world road surface conditions.

For model development, ten open-source crack segmentation datasets or subsets were used. These datasets were selected to include road pavement cracks, supplementary crack patterns from concrete or wall-like surfaces, and hard-negative non-crack samples. A brief description of each dataset is provided below.

The crack and non-crack public dataset was constructed from ten publicly available datasets to improve crack-pattern diversity and background generalization. Specifically, pavement crack samples were collected from Crack500 (3363 pairs), CrackTree200 (206 pairs), DeepCrack (521 pairs), Aidonchuk/Sylvie Chambon (185 pairs), GAPs384 (509 pairs), and the CrackForest Dataset (118 pairs), covering various asphalt textures, crack widths, and irregular crack structures. In addition, Rissbilder for Florian (3822 pairs), Volker (990 pairs), and Eugen Muller (55 pairs), which mainly contain concrete or wall-like surface cracks, were included to enhance thin-crack and background-texture diversity. Finally, 1411 non-crack samples were used as hard-negative data to reduce false crack detections. In total, 11,180 samples were used in this study. Representative examples of each dataset are shown in Figure 2a–j [13–22].

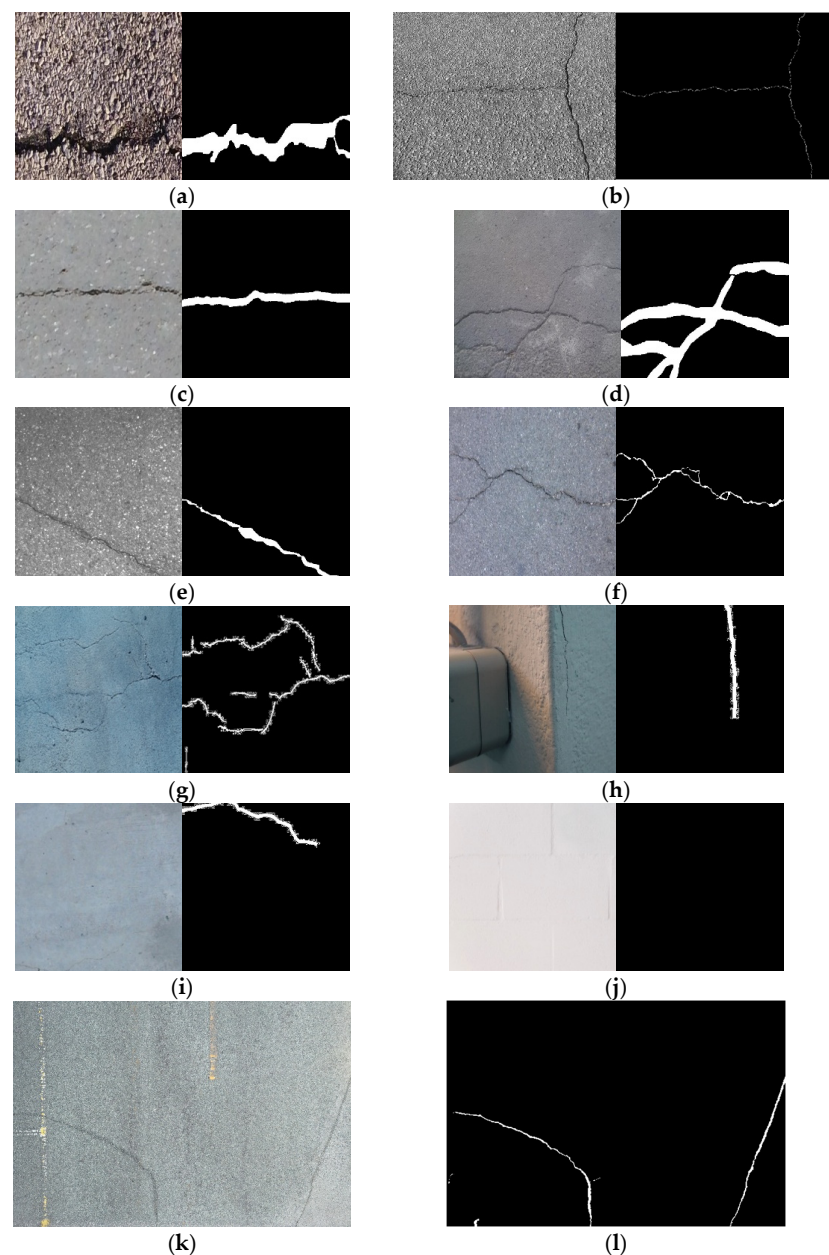


Figure 2. Representative examples of public crack datasets and UAV field imagery used in this study: (a) Crack500; (b) CrackTree200; (c) DeepCrack; (d) Sylvie Chambon; (e) GAPs384; (f) CrackForest Dataset; (g) Rissbilder for Florian; (h) Volker; (i) Eugen Muller; (j) Noncrack; (k) UAV field image; (l) Predicted binary mask for UAV field image.

Figure 2 shows representative examples of the public crack datasets and UAV field imagery used in this study. Figure 2a–j presents examples from the public datasets used for model development. For most public crack datasets, each panel shows an example image and its corresponding binary mask, whereas the Noncrack dataset contains non-crack surface samples. Figure 2k,l shows a representative UAV field image and its corresponding predicted binary mask, which illustrates the type of field imagery used for transferability evaluation.

For field transferability evaluation, UAV images were acquired from campus roads and parking areas at Pukyong National University. The survey area included internal campus roads where pedestrians and vehicles coexist, as well as various crack-like non-crack features such as lane markings, pavement joints, manholes, repaired pavement traces, shadows, surface contamination, and specular reflections. Therefore, the site provided practical field conditions for evaluating model behavior under crack-like non-crack road-surface features.

The UAV survey sections were selected to ensure that the road surface and cracks were clearly identifiable. Areas excessively occluded by trees, facilities, or vehicles were excluded or minimized. UAV image acquisition was conducted between 11:00 and 13:00 to maintain comparable solar illumination conditions while still including different shadow and cloud-cover conditions.

A DJI Mavic 2 Pro (DJI, Shenzhen, China) was used for UAV image acquisition. Images were acquired in a near-nadir viewing geometry, with the camera angle set close to 90° relative to the ground surface to minimize geometric distortion. UAV image acquisition was conducted over three survey dates under different illumination, cloud-cover, flight-altitude, and ground sampling distance conditions. Two surveys were conducted under clear-sky conditions with cloud-cover values of 1 and 2, whereas one survey was conducted under mostly cloudy conditions with a cloud-cover value of 7. Flight altitude and ground sampling distance ranged from 4.54 to 9.91 m and from 0.97 to 2.16 mm/pixel, respectively. From these acquisition campaigns, 148 road-surface images were selected for field inference and quantitative patch-level evaluation, including 50, 48, and 50 images from 14, 16, and 23 March, respectively. These images were selected to cover different road sections, shadow conditions, spatial resolutions, and crack-like non-crack surface features. The detailed acquisition and image-use conditions are summarized in Table 1.

Table 1. UAV image acquisition conditions under different illumination, weather, altitude, and ground sampling distance settings.

Date	Start Time	End Time	No. of Acquired Photos	No. of Used Photos	Flight Altitude (m)	GSD (mm/pix)	Weather	Visibility	Cloud Cover
14 March 2026	11:12:08	11:58:47	130	50	9.91	2.16	Clear	Over 20 km	2
16 March 2026	11:18:56	11:55:05	76	48	5.10	1.14	Mostly cloudy	17.8 km	7
23 March 2026	11:22:53	12:09:11	147	50	4.54	0.97	Clear	Over 20 km	1

2.3. Candidate Segmentation Models

To effectively detect irregular, narrow, and discontinuous road cracks in UAV imagery, nine semantic segmentation model configurations were selected as candidate models. The candidate models included DeepLabV3+, U-Net, U-Net++, Swin-Unet, SegNet, DeepCrack19, and MFPA-Net. For DeepLabV3+ and U-Net++, both ResNet50 and ResNet101 backbones were applied to examine the effect of feature extraction capacity on crack segmentation performance [23].

The U-Net family was included because its encoder–decoder architecture and skip connections are effective for restoring spatial information in thin and discontinuous crack regions [24]. U-Net++ was considered suitable for this task because its nested dense skip connections reduce the semantic gap between encoder and decoder features, thereby improving the reconstruction of fine crack structures [25]. DeepLabV3+ was included to evaluate the effect of multi-scale contextual feature extraction using atrous convolution and atrous spatial pyramid pooling (ASPP), which can improve segmentation performance by capturing contextual information at multiple receptive-field scales [26]. SegNet was included as a representative encoder–decoder model that restores spatial information using pooling indices from the encoder during upsampling [27]. Swin-Unet was adopted to examine the applicability of a transformer-based U-shaped architecture capable of capturing global contextual information [28]. DeepCrack19 and MFPA-Net were included to compare crack-specific architectures designed for thin linear object detection and multi-scale feature representation [29].

These candidate models were selected to compare different segmentation architectures under the same public dataset training and validation conditions before the final model was applied to UAV field imagery.

2.4. Dataset Preprocessing, Training, and Refinement

In total, 11,180 image–mask pairs were integrated from the ten public datasets. Because the datasets had different original image resolutions, all image–mask pairs were standardized to a target size of 448×448 pixels before model training. Image resizing was performed using bilinear interpolation, while binary mask resizing was performed using nearest-neighbor interpolation to preserve binary label values. After resizing, the images were converted into tensor format, and pixel values were scaled from the original 0–255 range to the 0–1 range.

The integrated public dataset was randomly divided into training and validation subsets at an approximate ratio of 8:2, resulting in 8944 image–mask pairs for training and 2236 image–mask pairs for validation. The split was performed at the image–mask pair level using a fixed random seed of 42 while preserving each image and its corresponding mask as a pair. The same training and validation subsets were used for all candidate model comparisons.

To evaluate the effect of training data augmentation, each candidate model was trained under two conditions: without augmentation and with augmentation. In the without-augmentation condition, only resizing and normalization were applied to the training images. In the augmentation condition, additional geometric and photometric transformations were applied to the public training subset. These included conventional geometric augmentation, ShiftScaleRotate, RandomShadow, RandomGamma, GaussNoise, JpegCompression, and CoarseDropout [30]. These transformations were used to improve robustness to variations in crack orientation, scale, illumination, image noise, compression artifacts, and partial occlusion. For the validation subset, only resizing and normalization were applied to ensure a consistent performance comparison.

To ensure a fair comparison among candidate models, all models were trained using the same training subset and evaluated using the same validation subset. The input image size was fixed at 448×448 pixels, and the same preprocessing and evaluation procedures were applied to all candidate models. The Adam optimizer was used for candidate model comparison, with an initial learning rate of 1×10^{-3} and a weight decay of 5×10^{-5} [31]. A CosineAnnealingLR scheduler was applied to promote stable convergence during training.

After candidate model comparison, the selected model was used for confidence-based dataset refinement. The model was applied to the public training subset, and the maximum prediction probability was calculated for each image. Image–mask pairs with a maximum prediction probability of 0.8 or higher were retained as confidence-filtered training samples. As a result, 3596 image–mask pairs were selected for final retraining. The validation subset was not used for confidence-based filtering and was kept unchanged for performance evaluation.

For final retraining, the selected model was trained using the confidence-filtered training samples under the hyperparameter settings summarized in Table 2. The model was trained for 100 epochs using the AdamW optimizer [32] and binary cross-entropy with logits loss. A CosineAnnealingLR scheduler was used with a minimum learning rate of 1×10^{-6} . All model training and UAV inference experiments were conducted on an Ubuntu 22.04 workstation using Python 3.10.18, PyTorch 2.4.0, CUDA 12.1, and four NVIDIA RTX 4080 GPUs. Final retraining used PyTorch DistributedDataParallel with a batch size of 16 per GPU and mixed-precision training. The final weights were selected based on validation IoU and convergence stability. The final retrained model was subsequently applied to UAV-acquired road-surface images for field transferability evaluation.

Table 2. Hyperparameters used for confidence-filtered final model retraining.

Parameter	Value	Parameter	Value
Train/Validation	8:2 random split	Epochs	100
Model	U-Net++ ResNet50	Initial learning rate	1×10^{-3}
Optimizer	AdamW	Weight decay	1×10^{-4}
Loss function	BCEWithLogitsLoss	Batch size	16/GPU

2.5. UAV Field Inference and GT Patch Construction

The final U-Net++ model with a ResNet50 backbone was applied to the UAV-acquired road-surface images to evaluate field transferability under real-world conditions. Because the original UAV images had a high resolution of 5472×3648 pixels, each UAV image was divided into a non-overlapping 8×8 grid for patch-level processing and evaluation. This procedure generated 64 patches per UAV image, and each patch had a size of 684×456 pixels. The patch-based approach was used to preserve local crack details while allowing high-resolution UAV images to be processed and inspected at a manageable scale.

For each UAV patch, the trained model generated a crack probability map, where each pixel value represented the predicted probability of belonging to a crack region. Predicted binary crack masks were produced by applying threshold values to the probability map. Overlay images were generated by superimposing the predicted binary masks on the corresponding UAV image patches.

In this study, model-generated binary masks are referred to as predicted binary masks and were not treated as ground truth unless manually corrected. From the 148 UAV images, the 8×8 patch construction procedure generated 9472 candidate patches. Among these candidate patches, 316 UAV patches were randomly selected using a fixed random seed to construct a manually corrected field GT subset.

The selected random subset included diverse field-condition categories, such as clear cracks, fine or low-contrast cracks, lane markings, shadows, surface contamination, repaired pavement boundaries, manholes, specular reflections, and hard-negative non-crack regions. Model-generated binary masks were used only as initial annotation candidates. Each candidate mask was manually corrected by visual inspection of the corresponding original UAV patch. All corrections were made by referring to the original UAV image patches rather than relying solely on the predicted masks. FP regions caused by non-crack

objects were removed, and missed crack regions were manually added. The final corrected masks were used as manually corrected field GT reference masks for supplementary quantitative evaluation.

During quantitative field evaluation, the model predictions for the selected UAV patches were compared with the manually corrected GT masks for pixel-wise metric calculation.

2.6. Evaluation Strategy and Metrics

The evaluation strategy consisted of four complementary parts: public dataset benchmark evaluation, confidence-based filtering evaluation, supplementary quantitative field evaluation using manually corrected UAV GT patches, and expert visual assessment for threshold sensitivity and error factor analysis.

First, candidate segmentation models were evaluated using the public dataset validation subset. This evaluation established benchmark performance under comparatively idealized dataset conditions and was used for model comparison. To assess the effect of training data augmentation, models trained with and without augmentation were evaluated under the same validation conditions.

Second, the effect of confidence-based filtering was evaluated using the selected model. The model trained with the original training subset and the model retrained with the confidence-filtered subset were compared using the same public validation subset. This evaluation was conducted to examine whether retaining high-confidence image-mask pairs improved validation performance under controlled public dataset conditions.

Third, a supplementary quantitative field evaluation was conducted using the manually corrected UAV field GT patch subset. The predictions for the selected UAV patches were compared with the manually corrected GT masks on a pixel-by-pixel basis. True-positive (TP), false-positive (FP), false-negative (FN), and true-negative (TN) pixels were aggregated across all annotated field patches. Based on these aggregated values, accuracy, precision, recall, Dice similarity coefficient (DSC), and intersection over union (IoU) were calculated. This field patch evaluation was used to provide supplementary quantitative evidence under UAV field conditions and was not intended to replace full-scale annotation of all 148 UAV images.

Fourth, expert visual assessment was conducted to examine practical interpretability under real-world UAV conditions. Five evaluators assessed threshold sensitivity using two visual criteria: crack thickness and crack connectivity. Inter-rater reliability among the five evaluators was assessed using the intraclass correlation coefficient (ICC). Because the evaluator scores were continuous 0–10 ratings and the final visual score was calculated from the mean score of five evaluators, a two-way random-effects absolute-agreement model, ICC(2,k), was used [33]. ICC values were calculated separately for crack thickness, crack connectivity, and the integrated score. Crack thickness was evaluated by determining whether the predicted cracks were excessively thick or too thin compared with the visible crack morphology. Crack connectivity was evaluated by determining whether the natural flow and continuity of cracks were preserved. Threshold values from 0.1 to 0.9 were applied at intervals of 0.1. The threshold scores were used to identify a practical threshold range under the tested UAV conditions.

For the final UAV detection results, expert visual assessment was also used to record major FP and FN factors. This expert visual assessment was used to evaluate practical field interpretability and error tendencies, but it was not treated as a substitute for a dense pixel-level ground-truth-based accuracy assessment. The evaluation metrics used in this study are defined in Equations (1)–(5):

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN} \quad (1)$$

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

$$DSC = \frac{2TP}{2TP + FP + FN} \quad (4)$$

$$IoU = \frac{TP}{TP + FP + FN} \quad (5)$$

where TP, FP, FN, and TN denote true-positive, false-positive, false-negative, and true-negative pixels, respectively. Accuracy represents the proportion of correctly classified pixels among all pixels. Precision indicates the proportion of predicted crack pixels that are actual crack pixels, whereas recall indicates the proportion of actual crack pixels that are correctly detected. DSC and IoU evaluate the spatial overlap between the predicted crack regions and the ground-truth crack regions. Accuracy was used as an auxiliary metric because crack segmentation involves a strong class imbalance between crack and background pixels.

3. Results

3.1. Model Comparison and Final Model Performance

The candidate segmentation models were first evaluated under public dataset conditions to establish benchmark performance before field transferability testing. The evaluation was conducted using the integrated open-source crack datasets described in Section 2.2. Because these datasets mainly consisted of cropped or crack-centered pavement images in which crack regions were relatively clear, the validation results in this section were interpreted as benchmark performance under comparatively idealized dataset conditions. Table 3 summarizes the performance of the nine candidate segmentation model configurations trained with and without data augmentation. The training time represents the total time required to train each model configuration for 35 epochs and is reported in minutes.

Overall, the effect of data augmentation was model-dependent. Data augmentation did not consistently improve the public validation metrics, and several non-augmented models achieved higher DSC and IoU values than their augmented counterparts. Therefore, data augmentation was not interpreted as a universally beneficial factor for public validation performance in this study.

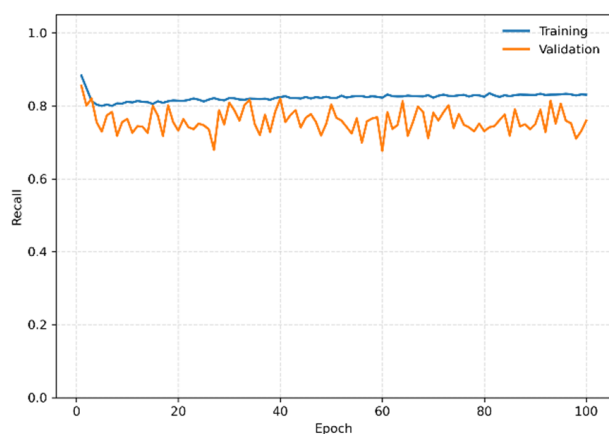
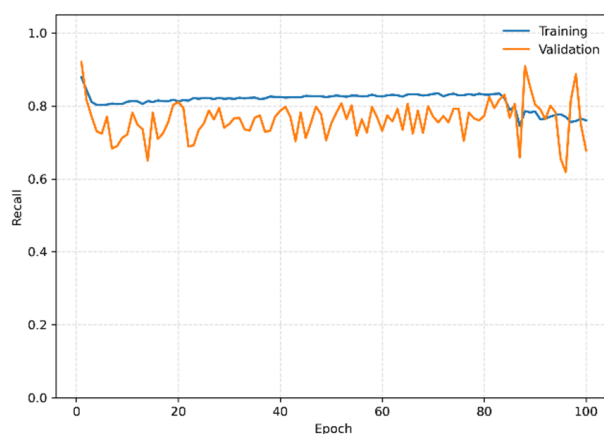
The best overlap-based performance on the public validation dataset was obtained by the non-augmented DeepLabV3+ models, particularly the ResNet50 backbone. However, the U-Net++ models showed comparable overlap-based performance and generally higher recall. This distinction was important because UAV-based crack inspection is intended to support field screening, where missed crack regions can be more problematic than limited over-detection. Thus, model selection was not based only on the highest DSC or IoU values.

DeepCrack19 and MFPA-Net without augmentation showed high precision but extremely low recall and overlap-based scores. These results indicate overly sparse crack predictions rather than balanced segmentation performance. Therefore, high precision alone was not considered sufficient for model selection, especially in a crack-screening application where failure to detect crack pixels can reduce practical usefulness.

Training time and validation stability were also considered. The U-Net++ ResNet50 model required less training time than U-Net++ ResNet101 while maintaining comparable segmentation performance. In addition, the validation recall curve was examined because recall stability is important for reducing missed crack regions during field screening. Figure 3 compares the training and validation recall curves of U-Net++ models with ResNet50 and ResNet101 backbones under the augmented training condition.

Table 3. Performance comparison of segmentation models on the public dataset.

Model	Augmentation	Precision	Recall	DSC	IoU	Training Time (min)
DeepCrack19	No	1.000	0.110	0.110	0.110	260
	Yes	1.000	0.110	0.110	0.110	272
DeepLabV3+ (ResNet50)	No	0.683	0.773	0.688	0.559	72
	Yes	0.636	0.709	0.601	0.477	73
DeepLabV3+ (ResNet101)	No	0.667	0.781	0.680	0.552	87
	Yes	0.553	0.797	0.595	0.463	90
MFPA-Net	No	0.823	0.111	0.111	0.111	108
	Yes	0.513	0.823	0.566	0.433	109
SegNet	No	0.589	0.802	0.619	0.488	67
	Yes	0.568	0.755	0.586	0.453	68
U-Net	No	0.653	0.805	0.672	0.543	63
	Yes	0.547	0.772	0.565	0.432	76
Swin-Unet	No	0.462	0.764	0.512	0.380	77
	Yes	0.395	0.687	0.419	0.300	79
U-Net++ (ResNet50)	No	0.635	0.842	0.680	0.551	104
	Yes	0.558	0.800	0.594	0.462	102
U-Net++ (ResNet101)	No	0.646	0.820	0.677	0.549	124
	Yes	0.519	0.801	0.564	0.434	127

**(a)****(b)****Figure 3.** Training and validation recall curves of U-Net++ models with ResNet50 and ResNet101 backbones: (a) U-Net++ with ResNet50; (b) U-Net++ with ResNet101.

As shown in Figure 3, the U-Net++ ResNet101 model showed a high recall value, but its validation recall curve showed relatively larger fluctuations during the later training stage. In contrast, the U-Net++ ResNet50 model showed more stable validation behavior while maintaining high recall. Considering recall, overlap-based performance, training time, and validation stability together, U-Net++ with a ResNet50 backbone was selected as the final model architecture. This selection is consistent with previous crack segmentation studies reporting that U-Net-based encoder–decoder architectures are effective for extracting thin and irregular crack structures because skip connections help preserve spatial details during decoder reconstruction [34,35].

Although the non-augmented models generally achieved better public validation DSC and IoU values, the final UAV application model was trained using the augmented

setting. This choice was not made because augmentation improved public validation performance. Rather, augmentation was retained as a field-robustness-oriented training setting to expose the selected model to broader geometric and photometric variations before UAV field inference. Because UAV field imagery contains more diverse acquisition and surface conditions than cropped public validation images, the augmented model was used for the final field application stage.

After selecting the U-Net++ ResNet50 architecture, confidence-based filtering was applied to the public training subset. Image-mask pairs with a prediction confidence of 0.8 or higher were retained, resulting in 3596 confidence-filtered training samples. Unlike Table 3, which reports the 35-epoch candidate model comparison results, Table 4 reports the performance of the selected U-Net++ ResNet50 architecture retrained under the final training setting for 100 epochs, with and without confidence-based filtering.

Table 4. Effect of confidence-based filtering on the selected U-Net++ ResNet50 model under the final 100-epoch training setting using the public validation dataset.

Model	Confidence Filtering	Precision	Recall	DSC	IoU
U-Net++ ResNet50	No	0.581	0.809	0.630	0.494
U-Net++ ResNet50	Yes, ≥ 0.8	0.802	0.787	0.794	0.659

As summarized in Table 4, confidence-based filtering improved precision, DSC, and IoU compared with the unfiltered augmented baseline, while recall slightly decreased. This indicates that the filtering process made the model more conservative but improved the spatial agreement of predicted crack regions with the validation masks. This trade-off is consistent with reliable-sample selection strategies in noisy-label learning, where reducing the influence of ambiguous or noisy samples can improve model robustness but may also reduce exposure to difficult samples [36]. Therefore, confidence-based filtering was interpreted as a refinement step that improved prediction reliability and overlap-based performance at the cost of a slight decrease in recall. The final confidence-filtered U-Net++ ResNet50 model was subsequently applied to the UAV field images.

3.2. UAV Field Inference and Supplementary Quantitative Patch Evaluation

The final U-Net++ model with a ResNet50 backbone was applied to 148 UAV images acquired from campus roads and parking areas. Unlike the public training and validation datasets, the UAV images contained various real-world road-surface factors, including lane markings, manholes, repaired pavement boundaries, shadows, surface discoloration or contamination, surface texture variations, and specular reflections. These factors visually resembled crack regions in some cases and therefore provided challenging field conditions for evaluating crack segmentation transferability.

As described in Section 2.5, UAV field inference was conducted using non-overlapping 8×8 patch-level processing. Figure 4 shows an example of the UAV-based inference result. The model detected major crack patterns with moderate or high visual contrast, and the overlay image confirmed that crack-prone regions could be identified in UAV-acquired road images. However, some non-crack regions were also detected as crack regions, and some fine or low-contrast cracks were not clearly extracted. These results indicate that the final model retained crack detection capability under UAV field conditions, but its predictions were affected by field-specific crack-like non-crack features.

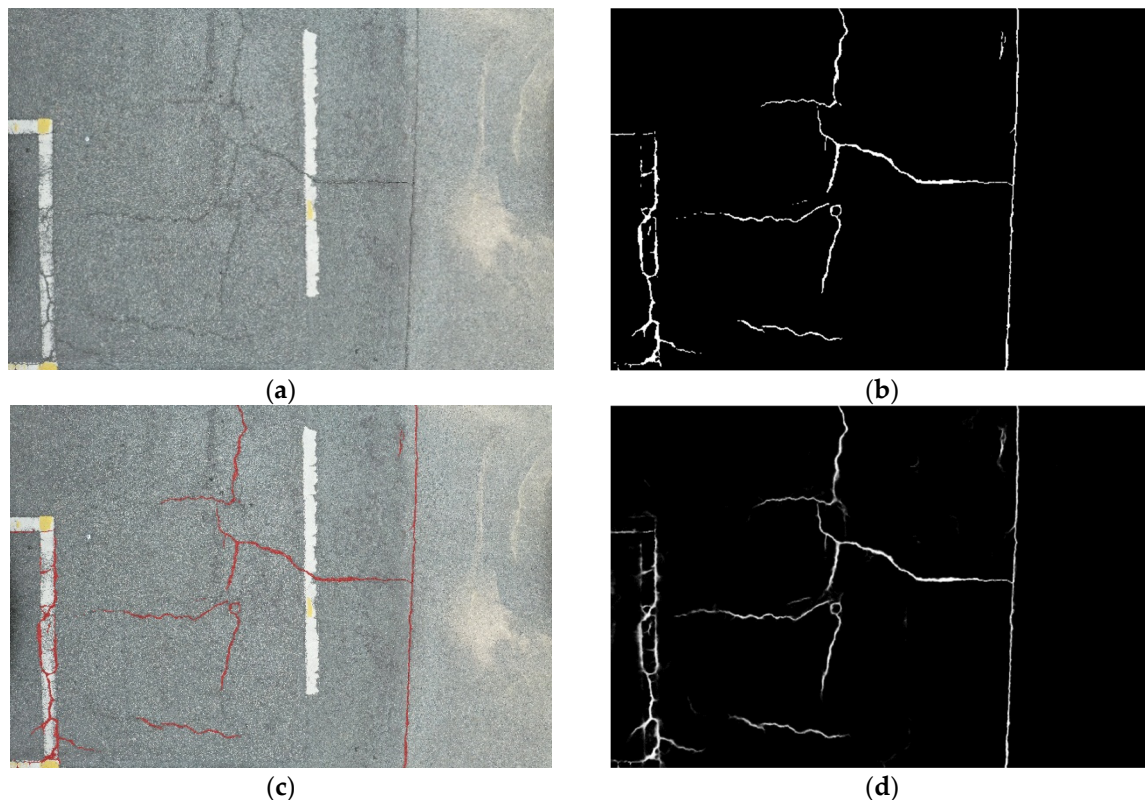


Figure 4. Example of UAV-based road crack detection results: (a) Original UAV image; (b) Predicted binary mask; (c) Overlay image; (d) Crack probability map.

In addition, the inference speed of the final U-Net++ ResNet50 model was measured using the 684×456 UAV patches used in the field evaluation. The model required an average of 84.31 ms per patch, corresponding to 11.86 patches/s. Considering that each original UAV image was divided into 64 non-overlapping patches, the estimated inference time was approximately 5.40 s per 5472×3648 UAV image.

To supplement the qualitative field interpretation, a quantitative patch-level evaluation was conducted using manually corrected UAV field GT reference masks. As described in Section 2.5, the UAV images were divided into 684×456 pixel patches, and 316 UAV patches were randomly selected from 9472 candidate patches using a fixed random seed. The selected patches included both crack-dominant regions and crack-like non-crack conditions, such as lane markings, shadows, manholes, repaired pavement boundaries, specular reflections, surface discoloration or contamination, and no-crack regions. Each selected patch contained 311,904 pixels, and the predicted mask and manually corrected GT mask were compared on a pixel-by-pixel basis. The proportions of TP, FP, FN, and TN pixels were calculated for each field-condition category.

Table 5 summarizes the pixel-wise confusion matrix results by field-condition category. The TN ratio was high in all categories, indicating that background pixels occupied a large proportion of the UAV field patches. Therefore, accuracy and other background-dominated measures should be interpreted carefully because high background classification performance does not necessarily indicate high crack segmentation performance. The Total row was calculated from the aggregated TP, FP, FN, and TN pixels across all 316 UAV field patches.

Table 5. Pixel-wise confusion matrix analysis between GT masks and predicted masks for UAV field patches. (Unit: %).

Category	No. of Patches	TP	FP	FN	TN
Fine Crack	81	2.58	0.18	0.26	96.98
No Crack	36	0.00	0.02	0.00	99.98
Cutting/Repaving Boundary	34	3.04	0.53	0.83	95.60
Specular Reflection	35	2.61	0.44	1.59	95.35
Shadow	35	3.79	0.33	0.55	95.33
Manhole	32	3.60	0.35	0.82	95.22
Surface Discoloration or Contamination	35	4.78	2.15	0.98	92.10
Paint	28	6.00	0.69	1.90	91.41
Total	316	3.12	0.53	0.75	95.60

Table 6 summarizes the category-wise quantitative evaluation results for the UAV field patches. Because the No Crack category did not contain GT crack pixels, crack-class metrics such as P, R, DSC, and IoU were not separately reported for this category. The Total row was calculated using the aggregated pixel-wise TP, FP, FN, and TN values across all 316 UAV field patches. Although the accuracy values were high across the categories, these values should be interpreted carefully because the UAV field patches contained a large proportion of background pixels. Therefore, P, R, DSC, and IoU were considered more informative than accuracy for evaluating crack segmentation performance.

Table 6. Quantitative evaluation results by UAV field patch category.

Category	Accuracy	P	R	DSC	IoU
Fine Crack	0.996	0.936	0.909	0.922	0.855
Cutting/Repaving Boundary	0.991	0.920	0.873	0.896	0.812
Specular Reflection	0.988	0.911	0.815	0.860	0.754
Shadow	0.986	0.851	0.786	0.817	0.691
Manhole	0.980	0.855	0.621	0.719	0.562
Surface Discoloration or Contamination	0.974	0.897	0.760	0.822	0.698
Paint	0.969	0.690	0.830	0.754	0.605
Total	0.987	0.856	0.806	0.830	0.710

The Fine Crack category showed the highest overall segmentation performance among the field-condition categories. This result indicates that visually distinguishable crack regions could be segmented effectively under the selected UAV field patch conditions. However, the Total performance was lower than the Fine Crack category alone, showing that the inclusion of crack-like non-crack field conditions reduced the overall field evaluation results.

Among the crack-like non-crack categories, Paint showed the lowest precision. This indicates that lane-marking paint and degraded painted regions tended to generate false-positive responses because their linear shapes, high local contrast, and irregular boundaries were visually similar to crack patterns. In contrast, Manhole showed the lowest recall, DSC, and IoU. This suggests that crack regions around manhole boundaries were more likely to be missed or fragmented, likely because circular manhole edges, local shadows, and boundary textures interfered with the continuity of crack predictions.

The Surface Discoloration or Contamination category showed a relatively high FP proportion in the confusion matrix results, indicating that local stains, discoloration, and contaminated pavement textures can also induce crack-like responses. These results confirm that real-world UAV imagery contains multiple non-crack surface features that can either

mimic crack patterns or obscure actual cracks. Therefore, evaluation using only clear crack or crack-dominant patches can overestimate field performance, whereas the inclusion of realistic crack-like non-crack conditions provides a more field-oriented estimate of UAV transferability.

3.3. Threshold Sensitivity Under Real-World Conditions

The model output was generated as a probability map, and a threshold was required to convert the probability map into a binary crack mask. To examine threshold sensitivity under real-world UAV conditions, nine candidate threshold values from 0.1 to 0.9 were applied at intervals of 0.1.

Figure 5 shows the predicted binary mask results generated under each threshold condition. At lower threshold values, such as 0.1 and 0.2, crack regions were detected more broadly, and crack connectivity was relatively well preserved. However, non-crack regions, including pavement texture, shadows, and parts of lane markings, were also detected as crack regions. At higher threshold values, such as 0.8 and 0.9, FP noise was reduced, but some fine or low-contrast crack segments were removed, and crack connectivity became weaker. These results indicate that threshold selection directly affects the balance between crack continuity and FP suppression.

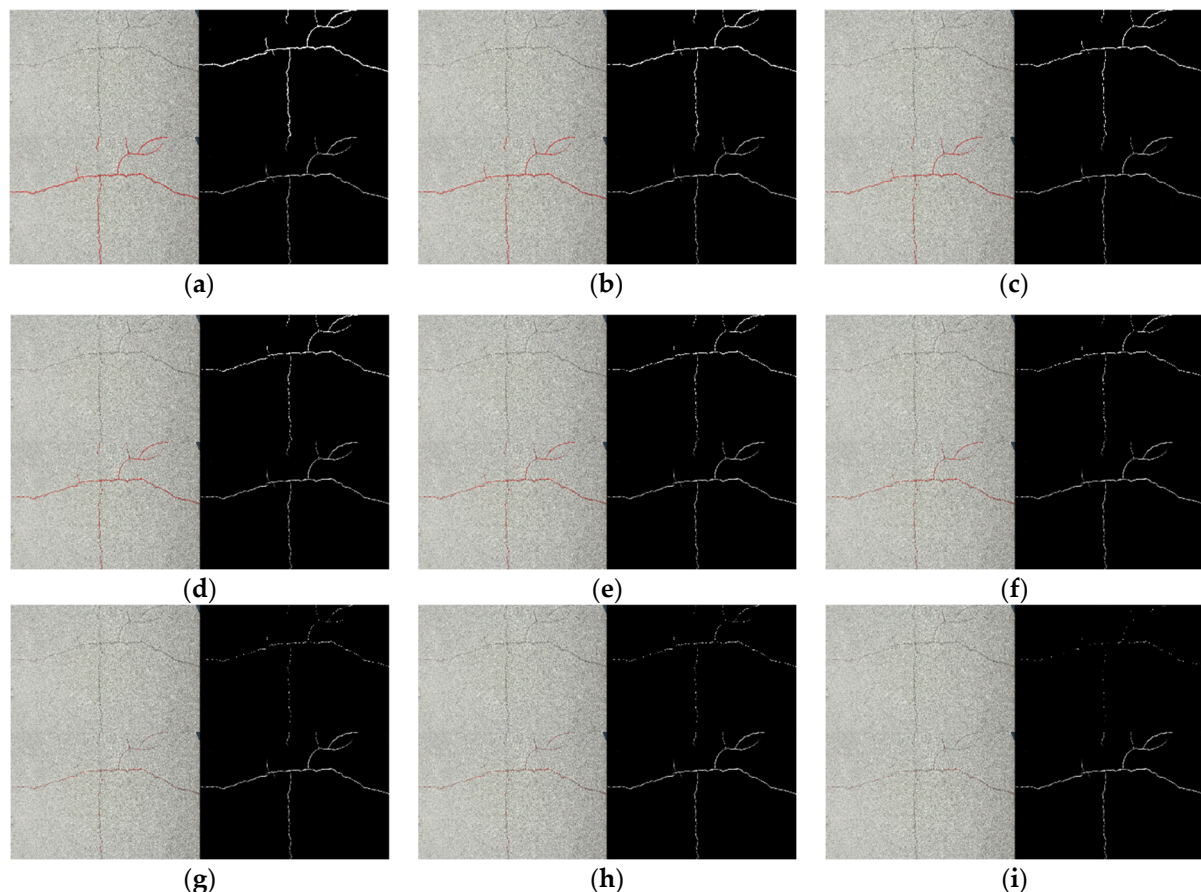


Figure 5. Predicted binary mask results according to threshold values from 0.1 to 0.9: (a) Threshold 0.1; (b) Threshold 0.2; (c) Threshold 0.3; (d) Threshold 0.4; (e) Threshold 0.5; (f) Threshold 0.6; (g) Threshold 0.7; (h) Threshold 0.8; (i) Threshold 0.9.

To support threshold selection from a practical visual interpretation perspective, five evaluators assessed the threshold results using two criteria: crack thickness and crack connectivity. Crack thickness was evaluated by determining whether the predicted cracks were excessively thick or too thin compared with the visible crack morphology in the

original UAV image. Crack connectivity was evaluated by determining whether the natural flow and continuity of cracks were preserved. Each criterion was scored on a 0–10 scale, where a higher score indicated better visual suitability.

As shown in Table 7, the crack thickness evaluation showed high scores in the threshold range of 0.5–0.6. Threshold 0.5 recorded the highest mean score of 7.76, and threshold 0.6 showed a similar mean score of 7.75. This result indicates that the predicted crack thickness was visually most appropriate in this threshold range. At lower threshold values, crack regions tended to be excessively expanded, whereas at higher threshold values, crack regions became too thin or partially disconnected.

Table 7. Evaluator scores for crack thickness and connectivity according to threshold values. (Thi: Thickness, Con: Connectivity).

Threshold	Evaluator A		Evaluator B		Evaluator C		Evaluator D		Evaluator E		Mean	
	Thi	Con	Thi	Con	Thi	Con	Thi	Con	Thi	Con	Thi	Con
0.1	2.21	5.44	1.88	5.45	0.83	6.44	2.77	5.11	2.17	5.88	1.97	5.66
0.2	3.47	7.41	3.30	7.85	1.88	6.60	3.99	7.11	3.21	7.30	3.17	7.25
0.3	5.04	8.17	5.12	8.34	2.92	6.17	5.28	8.27	4.59	7.81	4.59	7.75
0.4	6.91	7.58	6.92	7.30	4.40	5.40	7.03	7.69	6.46	7.56	6.34	7.11
0.5	8.21	6.00	8.15	5.89	5.88	4.26	8.43	6.14	8.13	6.06	7.76	5.67
0.6	7.94	4.21	8.06	4.12	6.85	3.41	7.75	4.48	8.13	4.26	7.75	4.10
0.7	6.13	3.17	6.34	3.03	6.32	2.48	5.59	3.21	6.59	3.13	6.19	3.00
0.8	4.02	2.02	4.10	2.01	4.90	1.66	3.16	2.03	4.73	2.00	4.18	1.94
0.9	1.06	1.00	1.14	1.00	3.26	0.83	1.01	1.01	1.00	1.00	1.49	0.97

The crack connectivity evaluation showed high scores in the threshold range of 0.3–0.4. Threshold 0.3 recorded the highest mean connectivity score of 7.75, followed by threshold 0.4 with a mean score of 7.11. The mean connectivity score decreased after threshold 0.5, indicating that low-confidence crack pixels were increasingly removed as the threshold increased. This result shows that higher threshold values can suppress FP responses but may also weaken the continuity of fine or low-contrast cracks.

The integrated evaluation results are shown in Table 8. The integrated score was calculated by averaging the crack thickness and crack connectivity scores. Thresholds of 0.4 and 0.5 recorded the highest integrated mean scores, both with a value of 6.72. This result indicates that the threshold range of 0.4–0.5 provided the most balanced visual result between crack thickness and crack connectivity.

Table 8. Integrated evaluator scores for crack thickness and connectivity according to threshold values.

Threshold	Evaluator A	Evaluator B	Evaluator C	Evaluator D	Evaluator E	Mean
0.1	3.83	3.67	3.63	3.94	4.02	3.82
0.2	5.44	5.57	4.24	5.55	5.25	5.21
0.3	6.60	6.73	4.54	6.77	6.20	6.17
0.4	7.24	7.11	4.90	7.36	7.01	6.72
0.5	7.11	7.02	5.07	7.29	7.10	6.72
0.6	6.08	6.09	5.13	6.11	6.20	5.92
0.7	4.65	4.69	4.40	4.40	4.86	4.60
0.8	3.02	3.06	3.28	2.60	3.37	3.07
0.9	1.03	1.07	2.04	1.01	1.00	1.23

To assess the consistency of the evaluator scores, inter-rater reliability was calculated using ICC(2,k). The ICC(2,k) values for crack thickness, crack connectivity, and the integrated score were 0.975, 0.988, and 0.976, respectively. These values indicate excellent agreement among the five evaluators and support the consistency of the visual threshold assessment.

Based on the integrated visual assessment, the threshold range of approximately 0.4–0.5 was regarded as a practical operating range that balanced crack thickness and crack connectivity. For subsequent UAV visualization and error-factor analysis, a threshold of 0.416 was used as the selected operating threshold within this practical range. This value should not be interpreted as a universal or fully optimized threshold for road crack segmentation. Instead, it should be understood as a field-dependent operating threshold influenced by UAV image resolution, pavement surface condition, model output distribution, and visual interpretation criteria.

Therefore, the threshold analysis in this study was used to examine sensitivity and practical interpretability under the tested UAV conditions, rather than to establish a generally applicable calibration value. The practical meaning and limitations of threshold sensitivity are further discussed in Section 4.3.

3.4. Error Factor Analysis Under Real-World UAV Conditions

The selected operating threshold of 0.416 was applied to the 148 UAV-acquired images to generate the final predicted binary masks and overlay images. Based on these results, qualitative field evaluation and error factor analysis were conducted to identify major error factors under real-world UAV conditions.

In the best case, the main crack morphology and connectivity were preserved, and FP regions were relatively limited. In the worst case, FP and FN detections occurred due to complex road-surface conditions, including manhole boundaries, lane markings, repaired or repaved pavement boundaries, shadows, and surface texture variations. These results show that the final UAV detection performance was strongly affected by crack-like non-crack features present in real-world road imagery.

For the final qualitative evaluation, a six-level rating criterion was applied to the 148 UAV-acquired images. Grade 1 was assigned to results in which crack thickness and connectivity were stably maintained with minimal FP and FN regions. For results rated grade 2 or lower, penalty factors were recorded. When multiple penalty factors occurred simultaneously in a single image, the contribution weights were divided among the relevant factors.

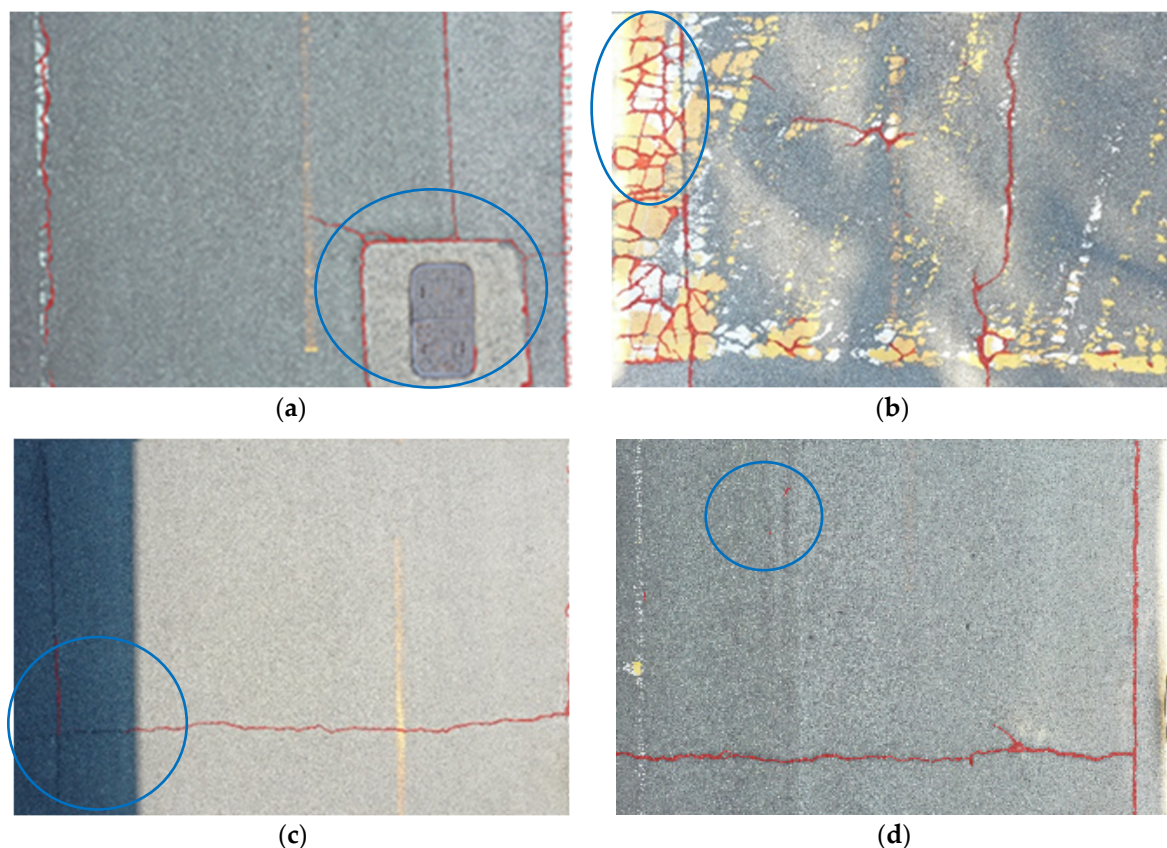
Table 9 summarizes the frequency and ranking of the major penalty factors observed in the final UAV detection results. Fine crack-like texture, FP and paint-related FP were the most frequent error factors, followed by shadow-induced FP, surface discoloration or contamination FP, cutting or repaving boundary FP, and manhole boundary FP. Specular reflection and object-induced FP occurred less frequently. These results indicate that the major error factors were not random noise but were strongly associated with specific crack-like non-crack surface features.

The penalty factor analysis showed that thin linear textures, degraded lane markings, manhole boundaries, repaired or repaved pavement boundaries, shadow boundaries, and surface contamination often produced visual patterns similar to cracks. These features generated FP detections because they shared crack-like characteristics, such as narrow linear shapes, dark intensity patterns, strong local contrast, or irregular boundaries. In addition, some fine or low-contrast cracks were partially missed when their visual contrast was weak or when their morphology was affected by surrounding surface conditions.

Table 9. Frequency and ranking of penalty factors in the final detection results using the selected operating threshold.

Penalty Factor	Frequency	Mean Rank	Mean Final Rank
Fine Crack FP	211	3.00	2.61
Paint FP	137	2.77	2.28
Shadow FP	78	3.65	2.07
Surface Discoloration/Contamination FP	72	2.71	1.93
Cutting/Repaving Boundary FP	66	3.18	2.32
Manhole FP	59	3.42	2.03
Crack FP	16	2.81	2.28
Specular Reflection FP	12	3.50	2.50
Object FP	2	2.00	1.50

Figure 6 presents representative examples of major FP factors observed in the UAV detection results. The examples show that non-crack features with thin linear shapes, strong local contrast, irregular boundaries, or dark intensity patterns were sometimes classified as crack regions. These results confirm that real-world UAV imagery contains various crack-like non-crack features that can reduce field detection reliability, even when the model shows acceptable benchmark performance on public crack datasets.

**Figure 6.** Representative examples of major False Positive (FP) factors observed in UAV detection results: (a) Fine Crack FP at the manhole boundaries; (b) Paint FP at the lane-marking paint; (c) Shadow FP in Shadow; (d) Surface Discoloration/Contamination FP.

4. Discussion

4.1. Benchmark Performance and UAV Field Transferability

Public datasets remain useful for model development, but the present results show that benchmark performance should not be equated with field performance. The final U-Net++ ResNet50 model achieved $P = 0.802$, $R = 0.787$, and $IoU = 0.659$ on public validation data, indicating that it learned public-dataset crack patterns. However, UAV field predictions were degraded by crack-like non-crack features.

The key contrast appeared when field categories were separated. The Fine Crack category produced high P/R/DSC values of 0.936/0.909/0.922, which were comparable to previous CrackForest-based studies reporting high values above 0.91 [15,34,37,38].

After field-specific non-crack elements were included, the total P/R/DSC decreased to 0.856/0.806/0.830. Paint and Manhole categories were especially problematic. This confirms the main finding of this study: evaluation on ideal or crack-only data can overestimate real-world UAV performance.

This discrepancy reflects the domain gap between public crack datasets and UAV imagery. Public datasets are often cropped or crack-centered with limited distractors, whereas UAV scenes include broader context, illumination variation, lane markings, shadows, contamination, repaired boundaries, manholes, specular reflections, and complex pavement textures that either mimic cracks or obscure actual cracks.

These findings agree with recent UAV pavement studies showing that cross-domain, cross-resolution, stitching, occlusion, illumination, and acquisition-condition differences can strongly affect crack segmentation in field applications [8,10]. Thus, field transferability testing is essential before deploying public-dataset-trained models.

4.2. Error Mechanisms Caused by Crack-like Non-Crack Features

The error factor analysis showed that many FP detections were caused by crack-like non-crack objects and surface conditions. Fine crack-like textures, lane-marking or paint regions, shadows, surface discoloration or contamination, repaired pavement boundaries, manhole boundaries, and specular reflections were identified as major penalty factors. These results suggest that the model responded strongly to visual features such as thin linear shapes, dark boundaries, local contrast changes, and irregular surface patterns.

This behavior indicates that the model may have learned low-level visual characteristics associated with cracks rather than fully distinguishing the semantic difference between cracks and non-crack road objects. For example, lane markings, repaired pavement boundaries, manhole edges, shadows, and surface contamination can all produce linear, curved, dark, or high-contrast patterns similar to cracks. As a result, these non-crack features were sometimes classified as crack regions.

FN cases can also be explained by similar mechanisms. Fine cracks, low-contrast cracks, cracks in shadowed areas, and cracks affected by specular reflection were sometimes missed because their visual contrast was weak or their morphology was partially obscured. This suggests that field errors are not caused only by model architecture but also by the mismatch between training data characteristics and real-world road-surface conditions. Previous studies have also noted that cracks are often narrow, discontinuous, and easily confused with pavement texture, shadows, stains, lane markings, or repaired surfaces, making field-level crack segmentation more difficult than benchmark validation [6,10].

The category-wise quantitative results also support this interpretation. The Fine Crack category showed a high DSC of 0.922, whereas categories containing visually similar non-crack structures showed lower performance. In particular, Paint showed the lowest precision, and Manhole showed the lowest recall, DSC, and IoU. This result indicates

that improving field robustness requires not only better crack extraction but also stronger discrimination between actual cracks and crack-like non-crack objects.

These findings suggest that future training datasets for UAV-based crack segmentation should include not only clear crack images but also diverse hard-negative samples, such as lane markings, manholes, repaired pavement boundaries, shadows, contamination, and specular reflections. This is consistent with previous comparative studies showing that the performance of semantic segmentation models can vary depending on crack morphology, image conditions, and evaluation metrics [39]. Therefore, field-oriented validation datasets are essential for evaluating whether models can distinguish actual cracks from visually similar non-crack road-surface features.

4.3. Practical Meaning of Threshold Sensitivity

The threshold sensitivity analysis demonstrated that the final crack detection result strongly depends on the threshold used to convert the crack probability map into a predicted binary mask. Lower thresholds preserved crack connectivity by including low-confidence crack pixels, but they also increased FP detections. In contrast, higher thresholds reduced noise and over-detection, but they removed fine or low-contrast crack segments and weakened crack connectivity.

In this study, crack thickness and crack connectivity showed different preferred threshold ranges. The crack thickness evaluation produced higher scores at thresholds of 0.5–0.6, whereas the crack connectivity evaluation produced higher scores at thresholds of 0.3–0.4. The integrated evaluation showed that thresholds of 0.4–0.5 provided the most balanced results. This range provided a practical balance between suppressing over-detection and preserving crack continuity.

However, the selected operating threshold of 0.416 should not be interpreted as a universal or fully optimized threshold for road crack segmentation. It was used as an operating value within the practical threshold range of 0.4–0.5 identified under the tested UAV conditions. This value can be influenced by UAV image resolution, flight conditions, pavement characteristics, model output distribution, and evaluator criteria. When the model is applied to other roads, flight altitudes, camera systems, pavement materials, seasons, or illumination conditions, the appropriate threshold may change. Therefore, threshold selection should be treated as a field-dependent sensitivity issue rather than as a fixed post-processing parameter.

This result also indicates that practical crack interpretation depends on the inspection objective. If the goal is to avoid missing fine cracks, a lower threshold may be useful because it preserves crack connectivity. If the goal is to reduce false alarms and minimize manual review, a higher threshold may be more appropriate. Therefore, UAV-based crack segmentation should allow adjustable threshold settings according to the intended maintenance or inspection purpose. Previous studies have also reported that thresholding can significantly affect crack extraction, noise suppression, and the separation of crack regions from background pavement areas in image-based crack detection [11,12].

4.4. Implications for UAV-Based Road Crack Inspection

The results of this study suggest that UAV-based crack segmentation can be useful for rapidly screening road-surface conditions over a relatively wide area. The final model was able to identify major crack-prone regions in UAV-acquired road images, and the overlay results provided interpretable visual outputs that can support field inspection. However, the results also show that fully automatic crack interpretation remains challenging under real-world UAV conditions because crack-like non-crack features can generate FP detections.

From a practical perspective, the model output should be interpreted as a decision-support layer rather than as a fully automated final diagnosis. Predicted binary masks and overlay images can help inspectors identify crack-prone regions efficiently, but human verification remains important when crack-like surface features, such as degraded lane markings, repaired pavement traces, manholes, shadows, or surface contamination, are present. Therefore, a human-in-the-loop inspection framework may be appropriate for UAV-based pavement monitoring, in which the model first highlights potential crack regions and inspectors then verify whether the detected regions correspond to actual cracks or crack-like non-crack objects. This approach can reduce manual inspection effort while maintaining reliability in complex field environments. UAV-based road damage studies have also emphasized the need to connect automated detection results with field verification and maintenance decision-making processes [40,41].

In addition, the results show that field-oriented evaluation should include both quantitative patch-level accuracy assessment and qualitative error factor analysis. Quantitative metrics provide objective evidence of segmentation performance, whereas error factor analysis reveals why the model fails under specific field conditions. Combining these two approaches is important for understanding not only how accurately a model performs, but also how reliably it can be used in practical UAV-based road inspection workflows.

These findings indicate that practical UAV-based crack inspection should combine quantitative model outputs with human verification. Future systems should further integrate field-oriented hard-negative training, targeted post-processing, adjustable thresholding, and human-in-the-loop review to improve reliability under complex road-surface conditions.

5. Conclusions

This study evaluated the field transferability of a public-dataset-trained road crack segmentation model using UAV-acquired road imagery. A U-Net++ model with a ResNet50 encoder backbone was selected through candidate model comparison and was further trained using confidence-based dataset refinement and training data augmentation. The final confidence-filtered model achieved P, R, DSC, and IoU values of 0.802, 0.787, 0.794, and 0.659, respectively, on the public validation dataset. The model was then applied to high-resolution UAV road-surface images using patch-level processing based on a non-overlapping 8×8 grid.

The UAV field results showed that visually distinguishable cracks could be segmented effectively, whereas crack-like non-crack factors reduced overall field reliability. In particular, lane-marking paint, manhole boundaries, repaired pavement traces, shadows, surface contamination, and specular reflections caused FP and FN detections. These results confirm that real-world UAV imagery contains field-specific distractors that are limited in many public crack datasets.

Threshold sensitivity analysis showed that the balance between crack connectivity and FP suppression strongly depends on the operating threshold. Under the tested UAV conditions, the threshold range of approximately 0.4–0.5 provided a practical visual balance between crack thickness and crack connectivity. The selected threshold of 0.416 was used only as an operating value within this range and should not be interpreted as a universal or fully optimized threshold.

Several limitations remain. The UAV images were acquired from a single campus site, and dense pixel-level GT masks were not constructed for all 148 UAV images. Instead, a supplementary quantitative evaluation was conducted using a manually corrected UAV field GT patch subset. Further validation is required under different road types, pavement materials, flight altitudes, camera systems, seasons, and illumination conditions. Future

work should expand UAV field datasets, include more hard-negative samples, and develop post-processing and human-in-the-loop verification strategies to reduce FP detections caused by crack-like non-crack objects.

Author Contributions: Conceptualization, S.L. and H.A.; methodology, S.L. and C.C.; software, H.K. and D.K.; validation, H.A. and C.C.; formal analysis, S.L. and C.C.; investigation, D.K. and H.K.; resources, J.J.; data curation, S.L. and D.K.; writing—original draft preparation, S.L. and C.C.; writing—review and editing, S.L.; visualization, H.K. and J.J.; supervision, H.A.; project administration, H.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The original contributions presented in this study are included in the article. Further inquiries can be directed to the corresponding author.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Haider, S.W.; Masud, M.M. Use of LTPP SMP Data to Quantify Moisture Impacts on Fatigue Cracking in Flexible Pavements [Summary Report]. 2020. Available online: <https://rosap.ntl.bts.gov/view/dot/43816> (accessed on 2 July 2026).
- Min, W.; Lu, P.; Liu, S.; Wang, H. A Review of Crack Sealing Technologies for Asphalt Pavement: Materials, Failure Mechanisms, and Detection Methods. *Coatings* **2025**, *15*, 836. [\[CrossRef\]](#)
- Yildirim, Y.; Qatan, A.; Prozzi, J. *Field Manual for Crack Sealing in Asphalt Pavements*; Texas Department of Transportation, Research and Technology Implementation Office: Austin, TX, USA, 2006.
- Choi, J.; Seo, J. Development of Budget Forecasting Procedure for Local Road Asphalt Pavement Maintenance Using LrPCI. *J. Korean Asph. Inst.* **2019**, *9*, 58–74. [\[CrossRef\]](#)
- Kim, S.H.; Park, J.T.; Kim, K.Y.; Shin, J.C.; Kim, J.R.; Baek, Y.J.; Lee, M.Y. A Study on Visual Inspection Method for Pavement Condition Evaluation. In *Proceedings of the Korean Society of Road Engineers*; Korean Society of Road Engineers: Seoul, Republic of Korea, 2017; p. 52.
- Wu, Y.; Li, S.; Li, J.; Yu, Y.; Li, J.; Li, Y. Deep Learning in Crack Detection: A Comprehensive Scientometric Review. *J. Infrastruct. Intell. Resil.* **2025**, *4*, 100144. [\[CrossRef\]](#)
- Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers. In *Proceedings of the Advances in Neural Information Processing Systems*; Curran Associates, Inc.: Red Hook, NY, USA, 2021; Volume 34, pp. 12077–12090.
- Shan, J.; Jiang, W.; Huang, Y.; Yuan, D.; Liu, Y. Unmanned Aerial Vehicle (UAV)-Based Pavement Image Stitching Without Occlusion, Crack Semantic Segmentation, and Quantification. *IEEE Trans. Intell. Transp. Syst.* **2024**, *25*, 17038–17053. [\[CrossRef\]](#)
- Shan, J.; Jiang, W.; Feng, X. Bridging Cross-Domain and Cross-Resolution Gaps for UAV-Based Pavement Crack Segmentation. *Autom. Constr.* **2025**, *174*, 106141. [\[CrossRef\]](#)
- Taha, H.; El-Habrouk, H.; Bekheet, W.; El-Naghi, S.; Torki, M. Pixel-Level Pavement Crack Segmentation Using UAV Remote Sensing Images Based on the ConvNeXt-UPerNet. *Alex. Eng. J.* **2025**, *124*, 147–169. [\[CrossRef\]](#)
- Chen, C.; Seo, H.; Jun, C.; Zhao, Y. A Potential Crack Region Method to Detect Crack Using Image Processing of Multiple Thresholding. *SIViP* **2022**, *16*, 1673–1681. [\[CrossRef\]](#)
- Fan, R.; Bocus, M.J.; Zhu, Y.; Jiao, J.; Wang, L.; Ma, F.; Cheng, S.; Liu, M. Road Crack Detection Using Deep Convolutional Neural Network and Adaptive Thresholding. In *Proceedings of the 2019 IEEE Intelligent Vehicles Symposium (IV)*, Paris, France, 9–12 June 2019; pp. 474–479.
- Yang, F.; Zhang, L.; Yu, S.; Prokhorov, D.; Mei, X.; Ling, H. Feature Pyramid and Hierarchical Boosting Network for Pavement Crack Detection. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 1525–1535. [\[CrossRef\]](#)
- Zou, Q.; Cao, Y.; Li, Q.; Mao, Q.; Wang, S. CrackTree: Automatic Crack Detection from Pavement Images. *Pattern Recognit. Lett.* **2012**, *33*, 227–238. [\[CrossRef\]](#)
- Liu, Y.; Yao, J.; Lu, X.; Xie, R.; Li, L. DeepCrack: A Deep Hierarchical Feature Learning Architecture for Crack Segmentation. *Neurocomputing* **2019**, *338*, 139–153. [\[CrossRef\]](#)

16. Aidonchuk Cracks_Segmentation_Dataset. Available online: https://github.com/aidonchuk/cracks_segmentation_dataset (accessed on 2 July 2026).
17. Augustauskas, R.; Lipnickas, A. Improved Pixel-Level Pavement-Defect Segmentation Using a Deep Autoencoder. *Sensors* **2020**, *20*, 2557. [[CrossRef](#)] [[PubMed](#)]
18. Eisenbach, M.; Stricker, R.; Seichter, D.; Amende, K.; Debes, K.; Sesselmann, M.; Ebersbach, D.; Stoeckert, U.; Gross, H.-M. How to Get Pavement Distress Detection Ready for Deep Learning? A Systematic Approach. In Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN), Anchorage, AK, USA, 14–19 May 2017; pp. 2039–2047. [[CrossRef](#)]
19. Shi, Y.; Cui, L.; Qi, Z.; Meng, F.; Chen, Z. Automatic Road Crack Detection Using Random Structured Forests. *IEEE Trans. Intell. Transp. Syst.* **2016**, *17*, 3434–3445. [[CrossRef](#)]
20. Pak, M.; Kim, S. Crack Detection Using Fully Convolutional Network in Wall-Climbing Robot. In *Proceedings of the Advances in Computer Science and Ubiquitous Computing*; Park, J.J., Fong, S.J., Pan, Y., Sung, Y., Eds.; Springer: Singapore, 2021; pp. 267–272.
21. Ham, S.; Bae, S.; Kim, H.; Lee, I.; Lee, G.-P.; Kim, D. Training a semantic segmentation model for cracks in the concrete lining of tunnel. *J. Korean Tunn. Undergr. Space Assoc.* **2021**, *23*, 549–558. [[CrossRef](#)]
22. Dorafshan, S.; Thomas, R.J.; Maguire, M. SDNET2018: An Annotated Image Dataset for Non-Contact Concrete Crack Detection Using Deep Convolutional Neural Networks. *Data Brief* **2018**, *21*, 1664–1668. [[CrossRef](#)] [[PubMed](#)]
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [[CrossRef](#)]
24. Ronneberger, O.; Fischer, P.; Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2015; Volume 9351, pp. 234–241.
25. Zhou, Z.; Siddiquee, M.M.R.; Tajbakhsh, N.; Liang, J. UNet++: A Nested U-Net Architecture for Medical Image Segmentation. *IEEE Trans. Med. Imaging* **2020**, *39*, 1856–1867. [[CrossRef](#)] [[PubMed](#)]
26. Chen, L.-C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. In *Computer Vision—ECCV 2018*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2018; Volume 11211, pp. 801–818.
27. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A Deep Convolutional Encoder-Decoder Architecture for Image Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [[CrossRef](#)] [[PubMed](#)]
28. Cao, H.; Wang, Y.; Chen, J.; Jiang, D.; Zhang, X.; Tian, Q.; Wang, M. Swin-Unet: Unet-Like Pure Transformer for Medical Image Segmentation. In *Computer Vision—ECCV 2022 Workshops*; Lecture Notes in Computer Science; Springer: Cham, Switzerland, 2023; pp. 205–218.
29. Jiang, X.; Mao, S.; Li, M.; Liu, H.; Zhang, H. MFPA-Net: An Efficient Deep Learning Network for Automatic Ground Fissures Extraction in UAV Images of the Coal Mining Area. *Int. J. Appl. Earth Obs. Geoinf.* **2022**, *114*, 103039. [[CrossRef](#)]
30. Buslaev, A.; Iglovikov, V.I.; Khvedchenya, E.; Parinov, A.; Druzhinin, M.; Kalinin, A.A. Albumentations: Fast and Flexible Image Augmentations. *Information* **2020**, *11*, 125. [[CrossRef](#)]
31. Kingma, D.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2014**, arXiv:1412.6980.
32. Loshchilov, I.; Hutter, F. Decoupled Weight Decay Regularization. In Proceedings of the International Conference on Learning Representations, New Orleans, LA, USA, 6–9 May 2019.
33. Koo, T.K.; Li, M.Y. A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *J. Chiropr. Med.* **2016**, *15*, 155–163. [[CrossRef](#)] [[PubMed](#)]
34. Escalona, U.; Arce, F.; Zamora, E.; Sossa, H. Fully Convolutional Networks for Automatic Pavement Crack Segmentation. *Comput. Syst.* **2019**, *23*, 451–460. [[CrossRef](#)]
35. Zhang, Z.; He, Y.; Hu, D.; Jin, Q.; Zhou, M.; Liu, Z.; Chen, H.; Wang, H.; Xiang, X. Algorithm for Pixel-Level Concrete Pavement Crack Segmentation Based on an Improved U-Net Model. *Sci. Rep.* **2025**, *15*, 6553. [[CrossRef](#)] [[PubMed](#)]
36. Song, H.; Kim, M.; Park, D.; Shin, Y.; Lee, J.-G. Learning From Noisy Labels With Deep Neural Networks: A Survey. *IEEE Trans. Neural Netw. Learn. Syst.* **2023**, *34*, 8135–8153. [[CrossRef](#)] [[PubMed](#)]
37. Fan, Z.; Wu, Y.; Lu, J.; Li, W. Automatic Pavement Crack Detection Based on Structured Prediction with the Convolutional Neural Network. *arXiv* **2018**, arXiv:1802.02208. [[CrossRef](#)]
38. Lau, S.L.H.; Chong, E.K.P.; Yang, X.; Wang, X. Automated Pavement Crack Segmentation Using U-Net-Based Convolutional Neural Network. *IEEE Access* **2020**, *8*, 114892–114899. [[CrossRef](#)]
39. Shi, Z.; Jin, N.; Chen, D.; Ai, D. A Comparison Study of Semantic Segmentation Networks for Crack Detection in Construction Materials. *Constr. Build. Mater.* **2024**, *414*, 134950. [[CrossRef](#)]

40. Chen, X.; Liu, C.; Chen, L.; Zhu, X.; Zhang, Y.; Wang, C. A Pavement Crack Detection and Evaluation Framework for a UAV Inspection System Based on Deep Learning. *Appl. Sci.* **2024**, *14*, 1157. [[CrossRef](#)]
41. Zhang, Y.; Zuo, Z.; Xu, X.; Wu, J.; Zhu, J.; Zhang, H.; Wang, J.; Tian, Y. Road Damage Detection Using UAV Images Based on Multi-Level Attention Mechanism. *Autom. Constr.* **2022**, *144*, 104613. [[CrossRef](#)]

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