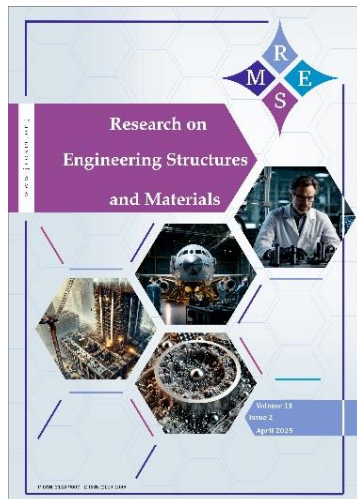




UAV remote sensing-based road corridor monitoring with spatial cross validation and deep semantic segmentation

Mustafa Hayri Kesikoglu, Soner Seker

Online Publication Date: 01 October 2026

URL: <http://www.jresm.org/archive/resm2026-1510ce0212rs.html>

DOI: <http://dx.doi.org/10.17515/resm2026-1510ce0212rs>

Journal Abbreviation: *Res. Eng. Struct. Mater.*

To cite this article

Kesikoglu M H, Seker S. UAV remote sensing-based road corridor monitoring with spatial cross validation and deep semantic segmentation. *Res. Eng. Struct. Mater.*, 2026; 12(5): 3141-3158.

Disclaimer

All the opinions and statements expressed in the papers are on the responsibility of author(s) and are not to be regarded as those of the journal of Research on Engineering Structures and Materials (RESM) organization or related parties. The publishers make no warranty, explicit or implied, or make any representation with respect to the contents of any article will be complete or accurate or up to date. The accuracy of any instructions, equations, or other information should be independently verified. The publisher and related parties shall not be liable for any loss, actions, claims, proceedings, demand or costs or damages whatsoever or howsoever caused arising directly or indirectly in connection with use of the information given in the journal or related means.



Published articles are freely available to users under the terms of Creative Commons Attribution - NonCommercial 4.0 International Public License, as currently displayed at [here](https://creativecommons.org/licenses/by-nc/4.0/) (the "CC BY - NC").

UAV remote sensing-based road corridor monitoring with spatial cross validation and deep semantic segmentation

Mustafa Hayri Kesikoglu ^{*,1,2,a}, Soner Seker ^{1,b}

¹Department of Civil Engineering, Faculty of Engineering and Natural Sciences, Uşak University, Uşak, Türkiye

²Department of Urban and Regional Planning, Faculty of Architecture and Design, Uşak University, Uşak, Türkiye

Article Info

Abstract

Article History:

Received 12 Feb 2026

Accepted 01 Oct 2026

Keywords:

UAV remote sensing;
Semantic segmentation;
Spatial cross validation;
Road corridor monitoring;
Spatial autocorrelation;
UAVid

UAV optical imagery provides the detail needed to map road surfaces, roadside vegetation, buildings and traffic participants in road corridors, but the accuracy of such maps is often estimated with random partitions of spatially and temporally autocorrelated frames. This study quantifies how much a random frame split inflates the accuracy of deep semantic segmentation compared with a sequence-based spatial cross validation that holds out complete UAV flight sequences. U-Net, DeepLabV3+ and SegFormer were trained under identical settings on all 42 labeled sequences (420 frames) of the UAVid benchmark of urban street scenes, together with a pixel-based Random Forest reference. Every frame was predicted once under each five-fold protocol, and the protocols were compared with paired sequence-level tests. The random split gave mIoU values of 72.9–74.0%, whereas sequence-based cross validation gave 68.6–69.9%. The difference of 4.1–4.3 percentage points was present in 41 or 42 of the 42 sequences for every model (Holm-adjusted Wilcoxon $p < 10^{-10}$), overall accuracy was inflated by 2.1–2.6 points, and the fold-to-fold variability of mIoU was two to three times larger under sequence-based validation. The inflation was largest for static cars (10.4–11.4 IoU points), moderate for road and low vegetation, small for buildings and trees and absent for pedestrians, while the ranking of the models did not change. Sequence-based validation therefore gives a more realistic estimate of the accuracy expected on new flights and is recommended for reporting UAV-based road corridor mapping results.

© 2026 MIM Research Group. All rights reserved.

1. Introduction

Road corridors are linear infrastructure systems whose condition depends on the pavement surface and on the land cover that surrounds it. Vegetation encroaching on the right-of-way (RoW), changes in adjacent built-up areas, parked and moving vehicles and the presence of pedestrians all influence operational safety and maintenance planning, and road agencies therefore need repeatable information on what occupies the corridor and how it changes over time [1]. Ground surveys provide this information at high thematic detail but are labor-intensive, expose field crews to traffic and cover only short sections per campaign. Satellite remote sensing offers systematic coverage of large areas, yet its revisit time, cloud contamination and spatial resolution limit the mapping of narrow corridor elements such as road verges, individual trees and vehicles [2–4].

Unmanned aerial vehicles (UAVs) fill this gap by acquiring optical imagery with centimeter-level ground sampling distance on demand and at low operational cost [5]. UAV image blocks are routinely processed with photogrammetric techniques into orthomosaics and digital surface

*Corresponding author: hayri.kesikoglu@usak.edu.tr

^aorcid.org/0000-0001-5199-0815; ^borcid.org/0000-0002-7632-9713

DOI: <http://dx.doi.org/10.17515/resm2026-1510ce0212rs>

Res. Eng. Struct. Mat. Vol. 12 Iss. 5 (2026) 3141-3158

models [6] and fused with terrestrial laser scanning point clouds for detailed 3D documentation [7], while the individual images and video frames are themselves a rich source for semantic interpretation. UAV studies have used such imagery to map vegetation communities, invasive species, land cover and hydrological variables with machine learning and deep learning [8–12], and to detect pavement damage along roads [13]. For road corridor monitoring, the classes of primary interest are the road surface, the adjacent buildings, roadside trees and low vegetation, and traffic participants such as vehicles and pedestrians. These classes are annotated in the UAVid benchmark, which provides densely labeled 4K frames of oblique UAV video over urban street scenes [14].

Deep semantic segmentation has become the standard approach for assigning a thematic label to every pixel of very high-resolution imagery. Fully convolutional networks [15] and encoder decoder architectures such as U-Net [16] established dense prediction with skip connections, DeepLabV3+ combined atrous convolution and spatial pyramid pooling with a light decoder [17], and SegFormer introduced a hierarchical transformer encoder with a simple multilayer perceptron decoder [18]. These model families are widely applied to remote sensing imagery [19] and to UAV scenes in particular [10, 20], and deep networks generally outperform pixel-based classical learners when spatial context matters [21].

The accuracy figures reported for such models are, however, only as informative as the protocol used to estimate them. Remote sensing data are spatially autocorrelated, so that observations close to each other tend to be more similar than distant ones [22]. When training and test samples are drawn at random from the same scenes, the test set contains samples that are near duplicates of training samples, and the estimated accuracy describes interpolation within the sampled area instead of prediction for new locations [23, 24]. Studies in ecology and remote sensing have shown that random cross validation can substantially overestimate map accuracy compared with spatially blocked validation [25, 26], and that the effect also applies to convolutional neural networks trained on UAV imagery [27]. Spatial and group-based resampling schemes, including block cross validation and leave-location-out or leave-one-flight-out designs, are therefore increasingly recommended [9, 12, 28–30], although the choice of validation strategy has to match the intended use of the map [31, 32].

UAV video data are particularly exposed to this problem, because frames extracted from the same flight share viewing geometry, illumination and scene content and often cover largely overlapping ground areas, so a random partition of frames places almost identical views of the same street into the training and the test set. Benchmark splits defined by dataset authors avoid this by assigning complete sequences to training, validation and test sets [14], but when UAV frames or patches are pooled and split at random, as is common in applied studies with a single collection of images, the size of the resulting optimism for corridor classes is rarely quantified. It is also unclear whether the optimism depends on the segmentation architecture or on the class, which matters when small and dynamic objects such as vehicles and pedestrians are part of the monitoring task.

This study addresses these questions with a controlled experiment on UAV optical imagery of urban road corridors. Three segmentation models that represent the main architectural families, U-Net, DeepLabV3+ and SegFormer, are trained under identical conditions on all 42 labeled sequences of UAVid, and a pixel-based Random Forest classifier with RGB spectral indices and local texture features serves as a simple classical reference. Accuracy is estimated with two five-fold cross validation protocols that differ only in how frames are assigned to folds: a random frame-level partition and a sequence-based partition in which all frames of a flight sequence are held out together. Because every frame is predicted exactly once under each protocol, the two estimates can be compared in a paired design at sequence level. The experiments quantify how much random splitting inflates overall and class-wise accuracy for corridor classes in urban street scenes, whether the inflation changes the ranking of the models and which classes remain difficult when a model is applied to unseen flights.

2. Materials and Methods

As summarized in Fig. 1, UAV optical frames and their reference labels were checked and down sampled, partitioned into folds with two cross validation protocols, used to train three deep

segmentation models and one classical reference under fixed settings, and assessed with pixel-based accuracy measures and sequence-level statistical tests.

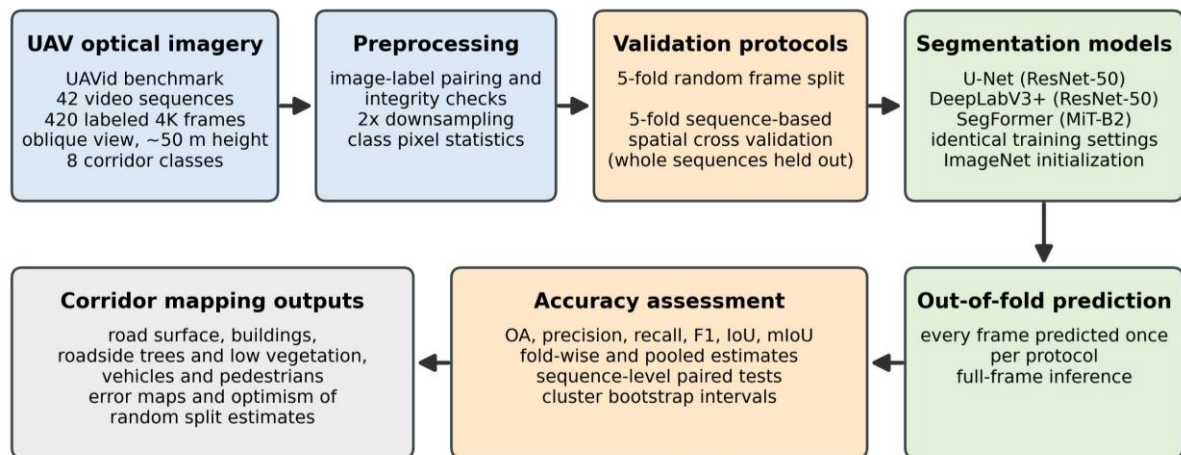


Fig. 1. Workflow of the study, from UAV optical imagery and validation design to accuracy assessment of corridor classes

2.1. Study Data and UAV Remote Sensing Context

The experiments use the UAVid semantic segmentation benchmark [14], which consists of oblique UAV video of urban and suburban street scenes (Table 1, Fig. 2). According to the dataset description, the imagery was recorded in 4K video mode at a flying height of approximately 50 m with the camera tilted by about 45° from the vertical, so that both the tops and the facades of objects are visible and the ground sampling distance increases from the near to the far part of each frame. Frame sizes are 3840 × 2160 or 4096 × 2160 pixels. Ten frames per sequence, taken at intervals of 100 video frames (about 5 s), were densely annotated with eight classes: building, road, tree, low vegetation, static car, moving car, human and background clutter. The original publication describes 30 sequences with 300 labeled images [14]; the dataset was later extended by 12 sequences, and the reference labels of the 15 test sequences were released by the dataset authors in 2023 through the UAVid website. All 42 sequences with 420 labeled frames were used in this study, and the original benchmark split was not used because both cross-validation protocols reassign all frames to folds.

Table 1. Summary of the data and validation designs used in this study

Item	Description
Dataset	UAVid, all 42 labeled sequences [14]
Platform and view	UAV video, oblique view (about 45°), flying height about 50 m
Labeled frames	420 (10 per sequence, about 5 s apart)
Original frame size (px)	3840 × 2160 (260 frames), 4096 × 2160 (160 frames)
Processed frame size (px)	1920 × 1080 or 2048 × 1080 (down sampling by 2)
Bands	RGB
Classes	building, road, tree, low vegetation, static car, moving car, human, clutter
Sources	RGB frames and train/val labels: DANS archive; test labels: third-party redistribution
Random split	5 folds of 84 frames assigned at random
Sequence-based CV	5 folds of 8 or 9 complete sequences (80 or 90 frames)

The data were obtained from two sources: the RGB frames of all 42 sequences and the color-coded reference labels of the 27 training and validation sequences were taken from the official UAVid archive in the DANS Data Station (doi:10.17026/dans-x9f-w9sa), and all 420 image files were verified against the SHA-1 checksums of this archive. The reference labels of the 15 test sequences are not contained in that archive; they were taken from a third-party redistribution on the Hugging

Face platform (dataset dronefreak/UAVid-2020), which provides the labels of all 42 sequences as class index images and states that the official labels were only reformatted. The index masks of this redistribution were used for all frames. For a random sample of 30 of the 270 training and validation frames they agreed in every pixel with the official color labels of the DANS archive. The test labels could not be compared with the official 2023 release, which is distributed through a separate file share that was not accessed in this study, and two anomalies found in them are described in Section 2.2. The dataset is distributed under a CC BY-NC-SA 4.0 license.

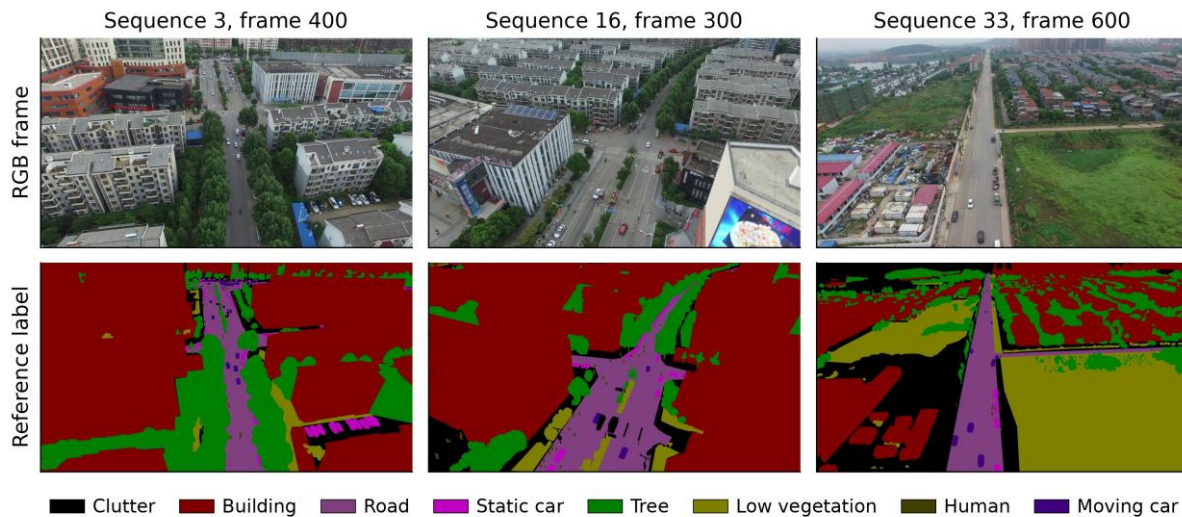


Fig. 2. Example UAVid frames from three sequences and the corresponding reference labels in the official color code

The UAVid classes cover the thematic content that is relevant for road corridor monitoring in urban areas. Road represents the carriageway and paved surfaces used by traffic, tree and low vegetation describe the roadside green that is subject to vegetation management within the RoW, building represents the adjacent built environment, static and moving cars represent parking and traffic, and human represents pedestrian presence. Background clutter collects all remaining surfaces and objects. The eight original classes were retained so that results remain comparable with the benchmark. The study is based entirely on RGB imagery; photogrammetric products such as orthomosaics or digital surface models were not derived, since the benchmark provides individual oblique frames without camera orientation data.

2.2. Data Preparation and Class Harmonization

Each image was paired with its label mask by sequence and frame index, and the pairs were checked for missing files, equal dimensions and label values outside the class range. One test frame (sequence 30, frame 600) contains 10,172 pixels (0.11% of the frame) without a valid class value; these pixels were excluded from training and evaluation. A plausibility check of the human class showed that the largest connected human region per frame is below 8,700 pixels in 99% of the frames, whereas two test frames (sequence 27, frame 300 and sequence 39, frame 300) contain human regions of 97,890 to 137,428 pixels covering lawn and paved areas that are labeled clutter, tree, low vegetation or road in the neighboring frames of the same sequences (Fig. A1). These regions (360,351 pixels in total) were treated as unlabeled in the evaluation. Because they were detected after training, they remained in the training data of the folds that contained the two frames, where they make up about 0.01% of the labeled pixels. The corrected evaluation is treated as the primary result, and the uncorrected evaluation is reported as a sensitivity analysis in Table A1; the masks of the excluded regions are available from the corresponding author upon reasonable request.

To make full-frame training and inference feasible on a 6 GB GPU, all frames were down sampled by a factor of two, using area interpolation for the RGB images and nearest neighbor interpolation for the masks. During training, square windows were sampled at random positions from the training frames at a random scale between 0.75 and 1.5 and resized to 512 × 512 pixels. Random

horizontal flipping and a random brightness and contrast change (gain 0.8 to 1.2, offset ± 20 gray values) were applied, and images were normalized with the ImageNet channel statistics. Vertical flips and rotations were not used, because they would produce viewing geometries that do not occur in oblique UAV imagery. Windows are drawn on the fly from the frames of the training folds only, so no patch is shared between training and test data.

2.3. Spatial Cross Validation Design

Two five-fold cross validation protocols were applied to the same 420 frames (Fig. 3). In the random protocol, frames were assigned to folds by a random permutation, independently of their sequence, which reproduces the common practice of splitting pooled UAV imagery at random. Because consecutive labeled frames of a sequence are only about 5 s apart and view largely overlapping parts of the same street, 386 of the 420 test frames (91.9%) in this protocol had at least one temporally adjacent frame of the same sequence in the training set, and on average 7.2 of the 9 other frames of their sequence were used for training.

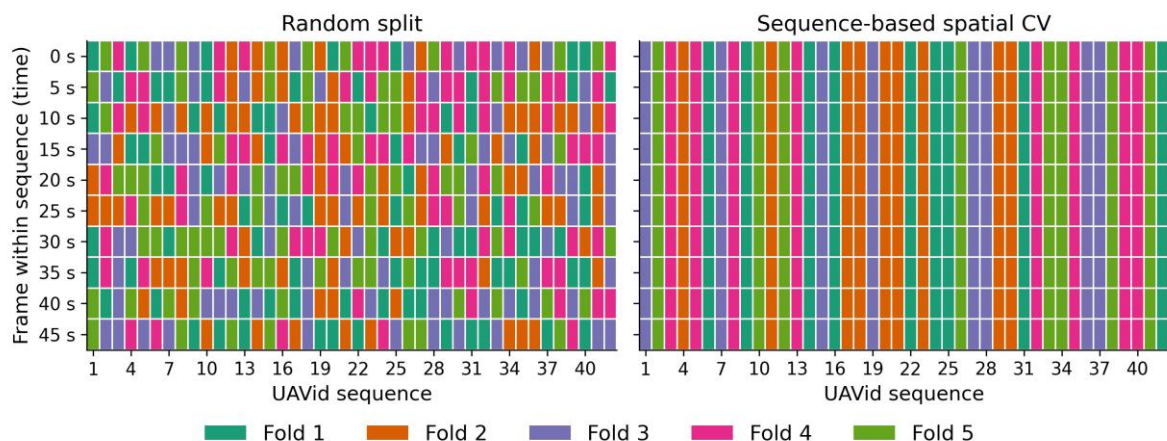


Fig. 3. Fold assignment of the 420 labeled frames under the random split (left) and the sequence-based spatial cross validation (right); columns are UAVid sequences and rows are the ten labeled frames of each sequence in temporal order

In the sequence-based protocol, the 42 sequences were randomly assigned to five groups of eight or nine sequences and all frames of a sequence were placed in the same fold, so that training and test data are sequence-disjoint. Since each sequence is a separate flight over a distinct street scene, this design approximates a leave-one-flight-out setting [23, 24, 28]. It does not guarantee geographic separation between folds, because location identifiers and flight campaign metadata are not provided with the dataset. The fold assignment of all frames is available from the corresponding author upon reasonable request.

In both protocols each fold was used once as the test set while the model was trained on the other four folds. No validation subset was used for early stopping or model selection within the cross validation: all models were trained for a fixed number of iterations and the final weights were evaluated. The only design decision informed by data of the study was the choice of the loss function, which followed short pilot runs on one sequence-based fold (Section 2.4). Every frame received exactly one out-of-fold prediction per model and protocol, which allows a paired comparison of the protocols.

2.4. Deep Semantic Segmentation Models

Three architectures that represent the main families of semantic segmentation networks were compared (Table 1). U-Net [16] is an encoder decoder network whose decoder recovers spatial detail by concatenating encoder features through skip connections; here it was built on a ResNet-50 encoder [33] with decoder stages of 256, 128, 64, 32 and 16 channels. DeepLabV3+ [17] applies atrous spatial pyramid pooling to encoder features at an output stride of 16 and fuses the result with low-level features in a light decoder; it was also built on ResNet-50. SegFormer [18] uses a hierarchical Mix Transformer (MiT) encoder with efficient self-attention and an all-MLP decoder

that aggregates features from four scales; the MiT-B2 encoder was used with a 256-channel decoder, because the 768-channel decoder of the original B2 configuration did not fit into GPU memory with the batch size used for the other models. The encoders were initialized with ImageNet pretrained weights [34] and the decoders randomly.

All models were trained with the same settings, and the loss was the sum of a class-weighted cross-entropy and a soft multiclass Dice loss [35]. Class weights were computed from the pixel frequencies f_c of the training folds as $w_c = (\text{median}(f)/f_c)^{1/2}$ and normalized to a mean of one, a damped form of median frequency balancing [36]. This loss was adopted as a limited preliminary choice to handle the class imbalance, after a pilot run on one sequence-based fold in which unweighted cross-entropy never predicted the human class. It was not the result of a systematic hyperparameter optimization: in short pilot runs of U-Net on the same fold, cross-entropy with Dice loss still failed to predict humans, whereas its class-weighted variant did, and this setting was retained without further tuning. The pilot runs are not part of the reported results, and the selected loss was applied unchanged to all models and both protocols. Because the pilot runs used data of the same study, the test folds cannot be regarded as entirely independent of this decision. Optimization used AdamW [37] with a learning rate of 1×10^{-4} and a weight decay of 0.01, a linear warm-up over 500 iterations followed by linear decay to zero, a batch size of eight crops and 6000 iterations. Mixed precision (16-bit floating point) arithmetic was used for training and inference. At test time each full frame was padded by reflection to a multiple of 32 pixels and predicted in one pass together with its horizontally flipped copy, whose logits were averaged. The random seed was fixed per fold, and the same seeds were used for both protocols.

Table 2. Segmentation models and computational cost; common settings are 512×512 crops, batch size 8, AdamW, learning rate 1×10^{-4} , weight decay 0.01, 6000 iterations and class-weighted cross-entropy plus Dice loss; training time is the median over the ten runs of each model and inference time the median over 24 full frames including flip averaging

Model	Encoder	Decoder	Parameters (M)	Training time per fold (min)	Inference (s per frame)
U-Net	ResNet-50	U-Net decoder (256–16 channels)	32.5	60	0.68
DeepLabV3+	ResNet-50, output stride 16	ASPP and decoder (256 channels)	26.7	48	0.55
SegFormer	MiT-B2	All-MLP decoder (256 channels)	24.7	74	2.08

As a simple classical reference, a pixel-based Random Forest classifier [38] with 100 trees and a minimum leaf size of five was trained on a limited set of features computed at a quarter of the original resolution: the RGB and HSV channels, the excess green index, the visible atmospherically resistant index and the normalized green red difference index, the local mean and standard deviation of brightness in 5×5 and 15×15 windows, the Sobel gradient magnitude and the relative row position in the frame. For each training frame up to 250 pixels per class were sampled, giving about 600,000 training pixels per fold. The reference was evaluated with the same folds and metrics as the deep models.

The experiments were implemented in Python with PyTorch 2.12, the Segmentation Models PyTorch library 0.5.0 and scikit-learn 1.8, and run on a laptop workstation with an NVIDIA GeForce RTX 2060 GPU (6 GB), an Intel Core i7-10750H CPU (2.60 GHz) and 32 GB RAM.

2.5. Evaluation Metrics

Accuracy was assessed from confusion matrices accumulated over all labeled pixels of the evaluated frames at the processed resolution [32, 39]. For class c , TP_c , FP_c and FN_c denote the numbers of true positive, false positive and false negative pixels, N is the total number of evaluated

pixels and $C = 8$ is the number of classes. Overall accuracy (OA), precision (P_c), recall (R_c), F1-score and intersection over union (IoU) are defined in Eqs. (1)–(5).

$$OA = \frac{\sum_{c=1}^C TP_c}{N} \quad (1)$$

$$P_c = \frac{TP_c}{TP_c + FP_c}, \quad R_c = \frac{TP_c}{TP_c + FN_c} \quad (2)$$

$$F1_c = \frac{2 P_c R_c}{P_c + R_c} \quad (3)$$

$$IoU_c = \frac{TP_c}{TP_c + FP_c + FN_c} \quad (4)$$

$$mIoU = \frac{1}{C} \sum_{c=1}^C IoU_c, \quad mF1 = \frac{1}{C} \sum_{c=1}^C F1_c \quad (5)$$

IoU is the Jaccard index used in the PASCAL VOC benchmark [40], and mIoU is the primary metric of the UAVid benchmark [14]. OA is reported for completeness, although it is dominated by the frequent classes. Two aggregation levels were used. Pooled estimates were computed from the sum of the confusion matrices of all 420 out-of-fold predictions, and fold-wise estimates were computed per test fold and summarized by their mean and standard deviation. In both cases every class occurs in the reference and in the predictions, so none of the denominators in Eqs. (2)–(4) is zero.

To test whether the protocols differ, the mIoU of each sequence was computed from the confusion matrix of its ten frames. Classes that occur neither in the reference nor in the prediction of a sequence have an undefined IoU and were omitted from the mean of that sequence. The paired sequence-level differences between the random and the sequence-based protocol were tested with the two-sided Wilcoxon signed-rank test [41], and 95% confidence intervals for the difference in pooled mIoU and OA were obtained with a cluster bootstrap that resamples the 42 sequences with replacement (10,000 replicates). The same sequence-level test was used to compare models within each protocol. The p -values were adjusted with the Holm procedure [42] within each family of comparisons.

3. Results

3.1. Dataset Characteristics and Class Distribution

The 420 frames contain 3.57×10^9 labeled pixels at the original resolution. Building (27.5%), tree (25.5%), clutter (18.9%), low vegetation (13.4%) and road (12.5%) account for 97.8% of the pixels, whereas static car (1.20%), moving car (0.89%) and human (0.12%) together account for 2.2% (Table 2, Fig. 4). The frequent classes are present in almost all frames and in all 42 sequences. Humans occur in 370 frames of 41 sequences, and moving cars in 334 frames of 39 sequences, so the rare classes are widespread but occupy small areas in each frame.

Table 3. Class distribution of the 420 labeled UAVid frames at the original resolution, before the correction of the two implausible human regions described in Section 2.2

Class	Pixels ($\times 10^6$)	Share (%)	Frames present	Sequences present
Building	982.8	27.51	419	42
Tree	911.2	25.51	420	42
Clutter	674.6	18.88	420	42
Low vegetation	479.4	13.42	420	42
Road	445.3	12.47	417	42
Static car	42.8	1.20	409	42
Moving car	31.9	0.89	334	39
Human	4.2	0.12	370	41

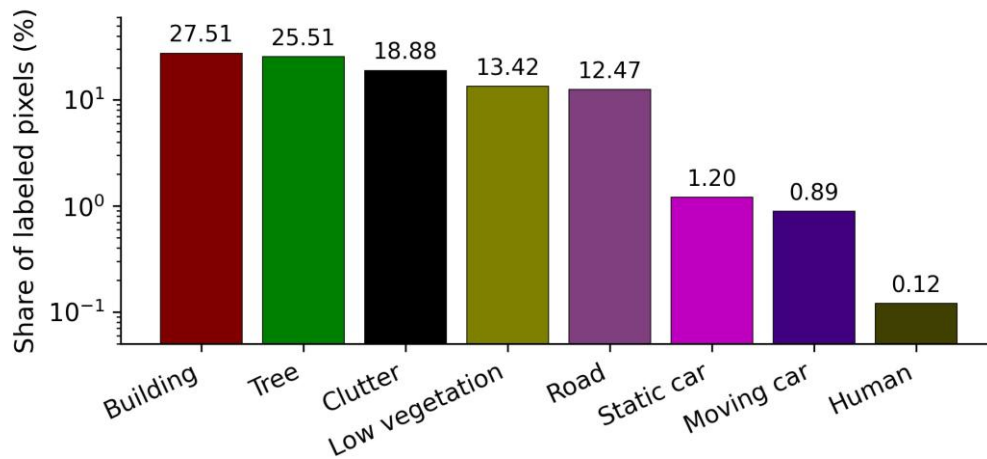


Fig. 4. Share of labeled pixels per class (logarithmic scale)

3.2. Model Performance under Random Split

Under the random frame split the three deep models reached mIoU values between 72.9% and 74.0% and OA values between 89.2% and 90.0% (Table 3, Fig. 5). U-Net obtained the highest mIoU (74.0%), followed by DeepLabV3+ (73.1%) and SegFormer (72.9%). The five folds gave very similar results, with standard deviations of the fold-wise mIoU between 0.7 and 0.9 points. The Random Forest reference reached 28.5% mIoU and 54.8% OA.

Table 4. Accuracy (%) under the five-fold random frame split; pooled values are computed from all 420 out-of-fold predictions

Model	OA	mIoU	mF1	Mean precision	Mean recall	Fold mIoU (mean $\pm$ SD)
U-Net	90.0	74.0	84.0	82.2	86.4	74.0 $\pm$ 0.9
DeepLabV3+	89.5	73.1	83.4	82.5	84.5	73.1 $\pm$ 0.7
SegFormer	89.2	72.9	83.4	81.9	85.4	72.9 $\pm$ 0.7
Random Forest	54.8	28.5	41.0	41.9	53.8	28.5 $\pm$ 0.5

The correction of the two implausible human regions (Section 2.2) raised the pooled mIoU of the deep models by 0.2 to 0.3 points and their human IoU by 1.8 to 2.4 points under both protocols, while the differences between the protocols changed by at most 0.1 point and all statistical conclusions remained unchanged (Table A1). All values reported below refer to the corrected evaluation.

3.3. Model Performance under Spatial Cross Validation

When complete sequences were held out, the mIoU of the deep models decreased to 68.6–69.9% and OA to 86.9–87.5% (Table 4). U-Net again obtained the highest mIoU (69.9%), followed by DeepLabV3+ (68.8%) and SegFormer (68.6%). The variation between folds increased markedly: the standard deviation of the fold-wise mIoU was 1.7 to 2.1 points, and fold-wise mIoU values ranged from 65.4% to 72.2%. The Random Forest reference reached 26.5% mIoU.

Table 5. Accuracy (%) under the five-fold sequence-based spatial cross validation

Model	OA	mIoU	mF1	Mean precision	Mean recall	Fold mIoU (mean $\pm$ SD)
U-Net	87.5	69.9	81.4	80.5	82.5	69.4 $\pm$ 2.1
DeepLabV3+	86.9	68.8	80.6	81.0	80.3	68.3 $\pm$ 2.1
SegFormer	87.1	68.6	80.5	79.7	81.5	68.2 $\pm$ 1.7
Random Forest	52.0	26.5	38.8	39.8	50.8	26.4 $\pm$ 1.0

The paired comparison of the two protocols is summarized in Table 5 and Fig. 5. For the deep models, the random split overestimated the pooled mIoU by 4.1 to 4.3 points, which corresponds to 5.5–5.9% of the random split estimate, and OA by 2.1 to 2.6 points. The 95% cluster bootstrap intervals of the mIoU difference did not include zero for any model. At sequence level, the random

split gave a higher mIoU for 41 of the 42 sequences with U-Net and for all 42 sequences with DeepLabV3+ and SegFormer; the median sequence-level difference was 3.3 to 4.0 points, and individual sequences showed differences of up to 16 points (Fig. 5c). The Random Forest reference showed a smaller absolute inflation of 2.0 mIoU points, but a similar relative inflation (7.2%).

Table 6. Difference between the random split (R) and the sequence-based spatial cross validation (S) in percentage points; confidence intervals from a cluster bootstrap over sequences and Holm-adjusted p-values from two-sided Wilcoxon signed-rank tests on sequence-level mIoU

Model	mIoU R	mIoU S	Δ mIoU (95% CI)	Relative Δ (%)	Δ OA (95% CI)	Sequences with lower mIoU	Adjusted p
U-Net	74.0	69.9	4.1 (3.0–5.3)	5.5	2.4 (2.0–2.9)	41/42	5.0×10^{-11}
DeepLabV3+	73.1	68.8	4.3 (3.4–5.3)	5.9	2.6 (2.1–3.1)	42/42	1.8×10^{-12}
SegFormer	72.9	68.6	4.3 (3.1–5.8)	5.9	2.1 (1.7–2.6)	42/42	1.8×10^{-12}
Random Forest	28.5	26.5	2.0 (1.5–2.6)	7.2	2.8 (2.2–3.6)	41/42	1.8×10^{-12}

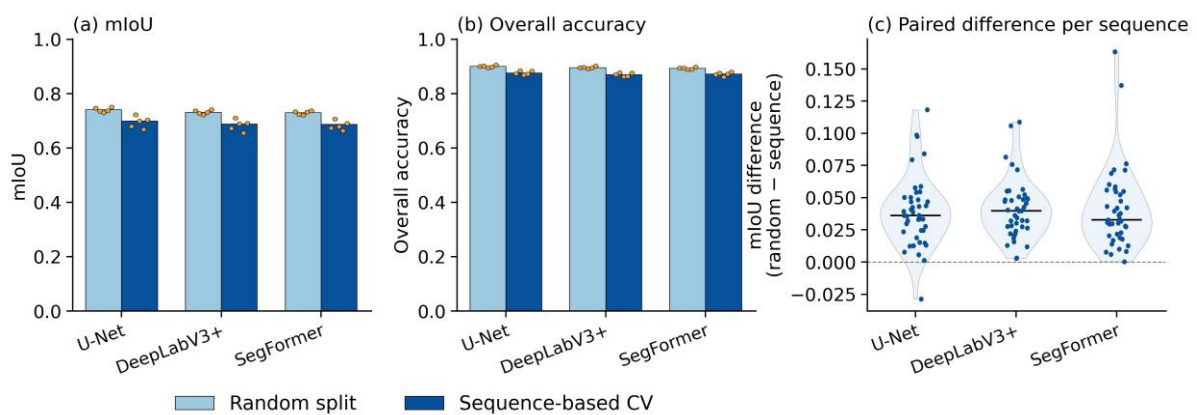


Fig. 5. (a) Pooled mIoU and (b) overall accuracy of the three models under both protocols, with dots showing the five fold-wise values; (c) paired difference in mIoU between the random split and the sequence-based cross validation for each of the 42 sequences, with horizontal lines marking the median

The ranking of the models was the same under both protocols, with U-Net achieving a higher sequence-level mIoU than DeepLabV3+ (39 of 42 sequences under both protocols) and SegFormer (37 and 34 of 42 sequences under the random and sequence-based protocols); the median differences were 1.0 to 1.7 points (Holm-adjusted $p < 0.001$). DeepLabV3+ and SegFormer did not differ significantly (adjusted $p = 0.20$ under the random split and $p = 0.80$ under the sequence-based validation). All deep models exceeded the Random Forest reference in every sequence, by at least 24 and in the median by 37 to 41 mIoU points.

3.4. Class Wise Mapping Accuracy and Qualitative Assessment

The class-wise IoU values in Table 6 and Fig. 6 show that, under sequence-based cross validation, buildings were mapped most accurately (IoU 89.3–90.7%), followed by road (80.1–80.8%) and tree (78.7–79.5%). Low vegetation (65.4–66.4%) and clutter (66.0–67.3%) reached intermediate values. Among the traffic classes, moving cars (68.1–72.4%) were mapped better than static cars (58.9–62.7%), and humans had the lowest IoU (38.8–39.8%).

The inflation caused by the random split differed strongly between classes. It was largest for static cars, with IoU differences of 10.4 to 11.4 points, and was driven by recall: the recall of static cars decreased from 86.3–88.6% under the random split to 72.2–75.4% under sequence-based validation, while precision changed by less than 3 points. Clutter (5.1–6.8 points), low vegetation (4.2–5.3 points), road (4.0–4.8 points) and moving cars (3.3–5.2 points) showed intermediate differences, and buildings (1.5–2.2 points) and trees (2.4–2.8 points) small differences. For

humans, the IoU under sequence-based validation was equal to or slightly higher than under the random split (−2.5 to +0.4 points).

Table 7. Class-wise IoU (%) under the random split (R) and the sequence-based spatial cross validation (S); mean Δ is the average difference R – S over the three deep models; class-wise F1-scores, precision and recall are available from the corresponding author upon reasonable request

Class	U-Net R	U-Net S	DeepLabV3+ R	DeepLabV3+ S	SegFormer R	SegFormer S	Mean Δ
Clutter	73.6	67.3	72.8	66.0	71.8	66.7	6.1
Building	92.3	90.7	91.5	89.3	91.2	89.8	1.7
Road	85.6	80.8	85.1	80.4	84.2	80.1	4.5
Static car	73.1	62.7	71.2	60.7	70.3	58.9	10.8
Tree	82.1	79.5	81.5	78.7	81.5	79.1	2.6
Low vegetation	71.4	66.1	70.6	65.4	70.6	66.4	4.9
Human	37.2	39.8	37.6	38.8	40.2	39.8	−1.1
Moving car	76.4	72.4	74.4	71.1	73.3	68.1	4.2

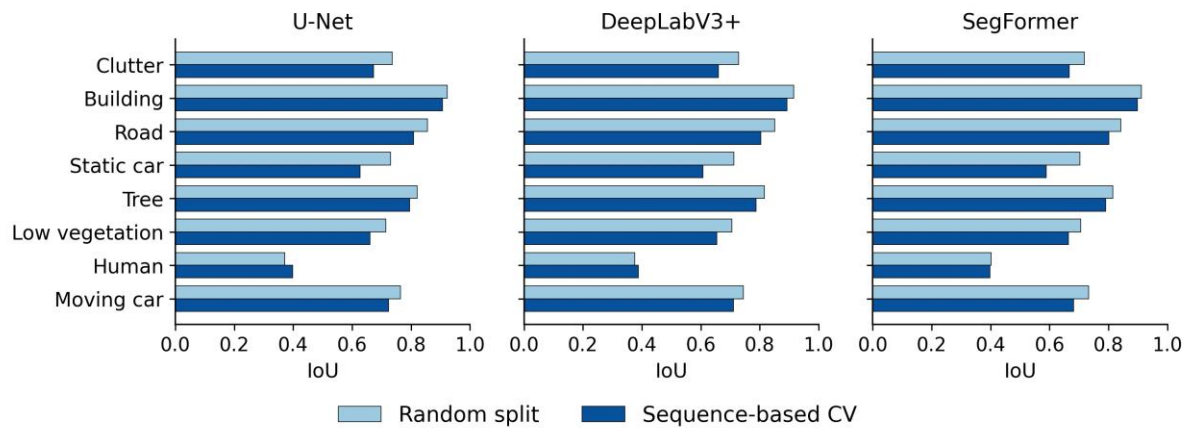


Fig. 6. Class-wise IoU of the three models under the random split and the sequence-based spatial cross validation

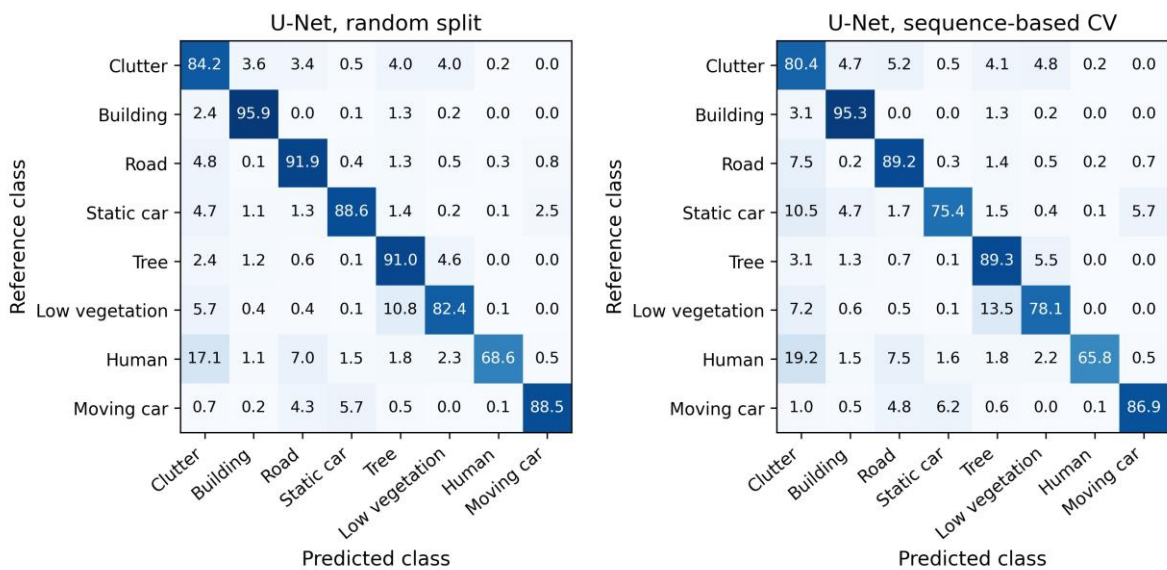


Fig. 7. Row-normalized confusion matrices (%) of U-Net under the random split and the sequence-based spatial cross validation

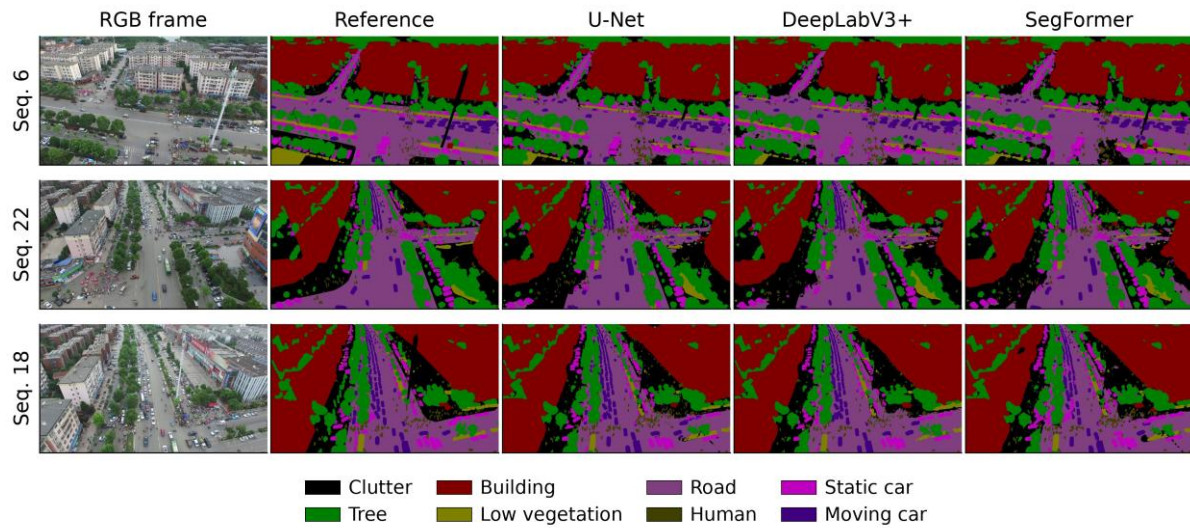


Fig. 8. Out-of-fold predictions under sequence-based spatial cross validation for frames of sequences 6, 22 and 18

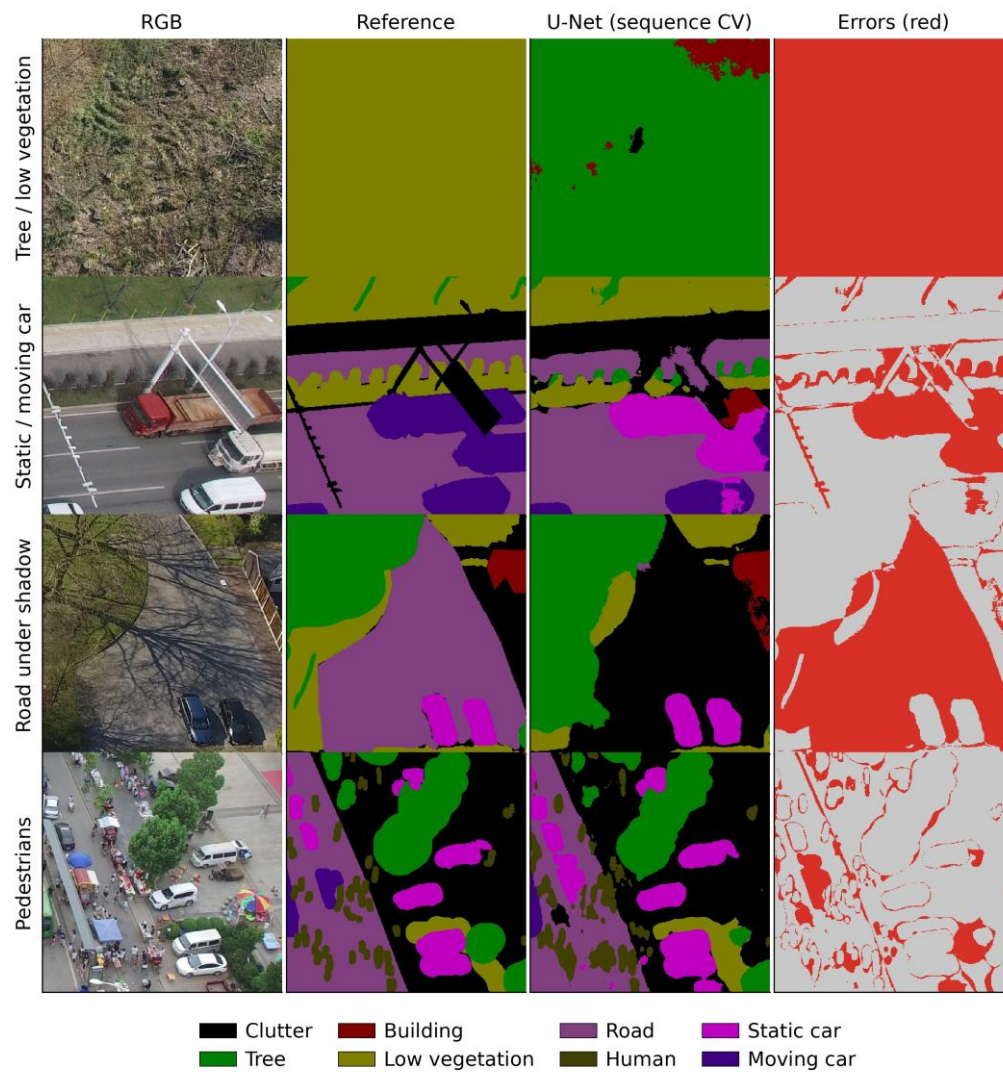


Fig. 9. Typical error situations of U-Net under sequence-based spatial cross validation: low vegetation labeled as tree, moving trucks labeled as static cars, road under tree shadow labeled as clutter and a crowded pedestrian area; crops of 360×360 pixels at the processed resolution, with red marking misclassified pixels

The confusion matrices of U-Net (Fig. 7) show where the additional errors on unseen sequences occur. Under sequence-based validation, 10.5% of the static car pixels were assigned to clutter and 5.7% to moving car, compared with 4.7% and 2.5% under the random split. Confusion between the two vegetation classes also increased: 13.5% of the low vegetation pixels were labeled as tree (10.8% under the random split), and 7.5% of the road pixels as clutter (4.8%). Humans were mainly confused with clutter (19.2%) and road (7.5%) under both protocols.

Fig. 8 shows out-of-fold predictions of the three models for frames of three sequences under sequence-based validation. All models reproduce the layout of the road network, building blocks, tree rows and parked vehicles, and their predictions differ mainly at object boundaries, in small vegetation patches and in the separation of static and moving cars. Typical error situations are shown in Fig. 9. Extended areas of low shrubs and cleared vegetation were classified as tree, large trucks on the carriageway were labeled as static instead of moving cars, road surfaces under strong tree shadows were assigned to clutter, and in crowded street scenes pedestrians were detected but with incomplete extents and confusion with clutter.

4. Discussion

4.1. Implications for Road Corridor Monitoring

The experiments indicate which corridor elements can be mapped reliably from RGB UAV imagery of urban streets when a model is applied to flights that were not used for training. Buildings, road surfaces and trees reached IoU values of about 80% or higher under sequence-based validation, which is sufficient for inventory tasks such as delineating the carriageway, identifying roadside tree rows or detecting changes in the adjacent built environment. Low vegetation and clutter were mapped with IoU values of about 66%, and static cars with about 60%, so maps of grass verges, shrubs and parked vehicles derived from new flights will contain a considerable share of misclassified pixels. Pedestrians remained the most difficult class. Their low IoU is consistent with the small number of pixels per person at a flying height of about 50 m, which was further reduced by the down sampling used in this study, and with the extreme class imbalance of the dataset (Table 2).

The sequence-based estimates are also of the same magnitude as benchmark results that were obtained on the sequence-disjoint official UAVid split. UNetFormer, for example, reported 67.8% mIoU on the UAVid test set [20], compared with 68.6–69.9% for the models of this study under sequence-based cross validation. Because that study used a different subset of sequences, image resolution and training schedule, the values are not directly comparable, but the agreement indicates that sequence-level separation yields accuracy estimates in the range known from benchmark evaluation, whereas the random split places the same models about four points higher. For practitioners who train models on their own corridor flights and evaluate them on randomly selected frames of the same flights, this optimism translates directly into overconfident expectations for subsequent survey campaigns.

4.2. Effect of Spatial Validation on Reported Performance

The consistent difference between the two protocols agrees with the literature on spatial autocorrelation in remote sensing and ecological mapping, which has repeatedly shown that random cross validation overestimates the accuracy of predictions for new locations [23, 25, 26]. Kattenborn et al. [27] reported the same effect for convolutional neural networks trained on UAV imagery of vegetation. The present results add to these findings in two respects. First, the inflation was observed in 41 or 42 of 42 sequences for every model and was of similar size for a CNN encoder decoder, an atrous pyramid network and a transformer, which suggests that it is a property of the data partition and not of a particular architecture. Second, random splitting also concealed the variability of the accuracy between groups of flights: the fold-wise standard deviation of mIoU increased from below one point to about two points under sequence-based validation, so a random split conveys an overly narrow picture of the uncertainty of the reported accuracy.

The class-wise pattern is consistent with the mechanism behind this optimism. In a random split, most test frames have neighboring frames from the same flight in the training set, taken about 5 s

earlier or later. Stationary objects such as parked cars, clutter objects and low vegetation patches appear in these neighboring frames at almost the same position and with the same appearance, so the model can in effect recognize objects it has already seen. This explains why static cars, clutter and low vegetation were the most inflated classes. Buildings and trees are large, frequent and visually distinctive, so they are mapped well even in unseen scenes and gained little from the leakage. Moving cars and pedestrians change position between frames, so the nearby training frames provide less direct information about their location, which is consistent with the smaller inflation for moving cars and the absence of inflation for humans. The magnitude of the overall effect is moderate compared with some large-area ecological mapping studies [25], which is plausible because all UAVid sequences were recorded with similar sensors, flying heights and urban scene types, so that unseen sequences are still similar to the training data in many respects.

The sequence-based design is not the only valid validation strategy, since spatial cross validation estimates the accuracy for new locations and can be pessimistic when the aim is to assess a map of the sampled area itself [31], and rigorous map accuracy assessment ideally relies on a probability sample of reference data [32]. For corridor monitoring, however, the typical use case is to apply a trained model to new flights, other road sections or later campaigns, and the sequence-based protocol matches this use case. The results therefore support grouping UAV frames by flight, sequence or location when accuracy is estimated, in line with the leave-location-out and leave-one-flight-out designs recommended for UAV mapping [9, 12] and with block cross validation in other spatial prediction tasks [28, 30].

4.3. Model Behavior across Corridor Classes

The three deep models behaved similarly across classes, and the choice of validation protocol did not alter their ranking. U-Net with a ResNet-50 encoder was about one point better than DeepLabV3+ and SegFormer, while the latter two did not differ significantly. This result should not be generalized to a statement on CNNs versus transformers: SegFormer was used with a reduced decoder width, all models were trained for the same fixed number of iterations on about 330 frames per fold, and larger transformer variants or longer schedules may benefit more from the data. The finding that the ranking was stable is nevertheless practically relevant, because it means that model comparisons made with a random split are not necessarily wrong, even though the absolute accuracies are inflated.

The main error sources were similar for all models, and the confusion between tree and low vegetation reflects the limited spectral separability of vegetation types in RGB imagery; in the oblique view, dense shrubs and low tree crowns have similar color and texture, and the distinction relies on height and context. The separation of static and moving cars is inherently difficult in single frames, as noted by the dataset authors [14], because the class depends on motion that is only observable over time; large trucks in traffic were typically labeled as static. Strong cast shadows on the road surface reduced the road recall, and pedestrians were missed or merged with clutter in crowded scenes. The pixel-based Random Forest reference, which used a limited set of RGB indices and local texture features at a quarter of the original resolution, reached less than 30% mIoU and almost failed for cars and humans. This shows that such a simple feature set is insufficient for oblique UAV street scenes, in line with earlier comparisons of deep and classical learners in remote sensing [8, 21], but it does not imply that classical classifiers with object-based or multi-scale features would perform equally poorly.

4.4. Limitations and Future Work

Several limitations should be considered when interpreting the results. The experiments are based on a single dataset of urban street scenes, and UAVid does not represent highway corridors with wider verges, embankments and cut slopes; the absolute accuracies and the size of the protocol effect may differ for rural and interurban road networks. The imagery is RGB only, so vegetation types and surface materials could only be separated by color and texture. Frames were down sampled by a factor of two, which reduces the detail available for small objects such as pedestrians, and all models were trained with a fixed budget of 6000 iterations and a common learning rate without model-specific tuning. Each protocol was evaluated with a single fold assignment; repeated

partitions with different random seeds were not computed, so the variability caused by the assignment itself is not quantified, although the consistency of the effect across 41 or 42 of 42 sequences indicates that the main result does not hinge on a particular assignment. The loss function was selected after short pilot runs on one sequence-based fold, so this fold was not entirely independent of the training design. The test labels were taken from a third-party redistribution, and two test frames contained implausible human labels, which were excluded from the evaluation but remained in the training data. Sequences were also treated as independent units, although some of them may have been recorded in neighboring parts of the same city; if so, the sequence-based estimates may still be slightly optimistic compared with a transfer to entirely new regions.

Future work should evaluate the transfer of models trained on UAVid to other UAV datasets with compatible classes, such as AeroScapes [43], and to corridor flights along highways, where validation by road section or flight campaign can be applied directly. Multispectral and thermal sensors could improve the separation of vegetation types and surface conditions, and temporal models that exploit consecutive video frames are a natural way to distinguish moving from static vehicles. Multi-temporal UAV and satellite observations would allow vegetation encroachment and land cover change along the RoW to be monitored over time [44], and combining corridor land cover maps with terrain information would support slope hazard assessment along road cuts and embankments [45]. For traffic participants, object detection with localization measures suited to tiny targets, such as the normalized Wasserstein distance [46, 47], could complement pixel-based segmentation. Estimating the area of applicability of a trained model [48] would further help to flag new flights whose conditions differ from the training data.

5. Conclusions

This study compared a random frame split with a sequence-based spatial cross validation for deep semantic segmentation of UAV optical imagery of urban road corridors. U-Net, DeepLabV3+ and SegFormer were trained under identical settings on all 42 labeled UAVid sequences, and every frame was predicted once under each protocol. Across the three models, the random split overestimated mIoU by 4.1 to 4.3 percentage points and overall accuracy by 2.1 to 2.6 points, and the optimism was present in 41 or 42 of the 42 sequences. It was concentrated in stationary and context-dependent classes such as static cars, clutter, low vegetation and road, whereas buildings and trees were affected little and pedestrians not at all. This class-wise pattern is consistent with information leakage between temporally adjacent frames of the same flight. Sequence-based validation also revealed a two to three times larger variability of accuracy between groups of flights, which a random split conceals.

On unseen flights, RGB-based deep segmentation mapped buildings, road surfaces and trees with IoU values of about 80% or higher, low vegetation and clutter with about 66%, vehicles with 59–72% and pedestrians with about 40%. These values indicate that the road surface, the adjacent built environment and roadside tree rows can be inventoried reliably from RGB UAV imagery of urban streets, whereas grass verges, shrubs, parked vehicles and pedestrians require further improvement before they can be mapped operationally. The relative ranking of the models was not affected by the validation protocol, with U-Net performing slightly better than DeepLabV3+ and SegFormer under the fixed training budget used here, and all deep models clearly outperformed a pixel-based Random Forest reference with a limited feature set.

For UAV-based road corridor monitoring, accuracy should be reported with validation schemes that hold out complete flights or sequences, because these estimates correspond to the intended application of trained models to new survey campaigns. Reporting class-wise metrics, the variability between folds and confidence intervals obtained at the level of flights or sequences, together with the fold assignment itself, makes such results easier to verify and compare. The transfer of these findings to highway corridors, to other sensors and to repeated partitions of the data remains to be tested in future studies.

Appendix A. Sensitivity to the Label Correction

Table A1. Main results with the uncorrected and the corrected evaluation of the two implausible human regions (%) under the random split (R) and the sequence-based spatial cross validation (S); Δ is R minus S mIoU with the 95% cluster bootstrap interval, sequences lower is the number of sequences with lower mIoU under S, and p is the unadjusted Wilcoxon signed-rank p-value

Evaluation	Model	mIoU R	mIoU S	Human IoU R	Human IoU S	Δ mIoU (95% CI)	Sequences lower	p
Uncorrected	U-Net	73.7	69.6	35.4	37.6	4.1 (3.1–5.3)	41/42	5.0×10^{-11}
Uncorrected	DeepLabV3+	72.8	68.5	35.6	36.4	4.3 (3.4–5.3)	42/42	4.5×10^{-13}
Uncorrected	SegFormer	72.6	68.3	38.1	37.7	4.3 (3.1–5.8)	42/42	4.5×10^{-13}
Uncorrected	Random Forest	28.5	26.5	1.3	1.2	2.0 (1.5–2.6)	41/42	9.1×10^{-13}
Corrected	U-Net	74.0	69.9	37.2	39.8	4.1 (3.0–5.3)	41/42	5.0×10^{-11}
Corrected	DeepLabV3+	73.1	68.8	37.6	38.8	4.3 (3.4–5.3)	42/42	4.5×10^{-13}
Corrected	SegFormer	72.9	68.6	40.2	39.8	4.3 (3.1–5.8)	42/42	4.5×10^{-13}
Corrected	Random Forest	28.5	26.5	1.3	1.2	2.0 (1.5–2.6)	41/42	9.1×10^{-13}

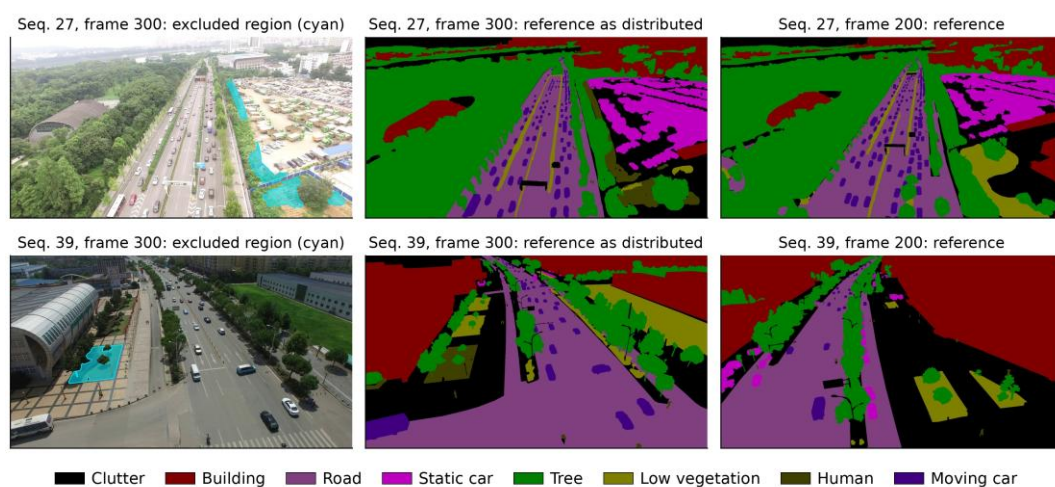


Fig. A1. Implausible human labels in two test frames: the excluded regions (cyan, left), the reference labels as distributed (middle) and the reference labels of the preceding labeled frame of the same sequence (right), in which the corresponding areas are labeled clutter, tree, low vegetation or road

Data Availability

The UAVid dataset is available from the UAVid website (<https://uavid.nl/>) and from the DANS Data Station (doi:10.17026/dans-x9f-w9sa) under a CC BY-NC-SA 4.0 license. The fold assignment, the corrected and uncorrected per-frame, per-sequence and class-wise metrics, the masks of the excluded label regions, the training and evaluation scripts and the software environment are available from the corresponding author upon reasonable request.

Acknowledgement

The authors thank the creators of the UAVid dataset for making the imagery and reference labels publicly available.

References

- [1] Outay F, Mengash HA, Adnan M. Applications of unmanned aerial vehicle (UAV) in road safety, traffic and highway infrastructure management: Recent advances and challenges. *Transportation Research Part A: Policy and Practice*, 2020; 141: 116 - 129. <https://doi.org/10.1016/j.tra.2020.09.018>
- [2] Zhu XX, Tuia D, Mou L, Xia GS, Zhang L, Xu F, Fraundorfer F. Deep Learning in Remote Sensing: A Comprehensive Review and List of Resources. *IEEE Geoscience and Remote Sensing Magazine*, 2017; 5(4): 8 - 36. <https://doi.org/10.1109/MGRS.2017.2762307>

- [3] Ma L, Liu Y, Zhang X, Ye Y, Yin G, Johnson BA. Deep learning in remote sensing applications: A meta-analysis and review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2019; 152: 166 - 177. <https://doi.org/10.1016/j.isprsjprs.2019.04.015>
- [4] Kesikoglu MH. Evaluating spatial resolution enhancement on impervious surface detection using support vector machine. *Research & Design*, 2024; 1(1): 33 - 42. <https://doi.org/10.17515/rede2024-003bu1111rs>
- [5] Colomina I, Molina P. Unmanned aerial systems for photogrammetry and remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2014; 92: 79 - 97. <https://doi.org/10.1016/j.isprsjprs.2014.02.013>
- [6] Nex F, Remondino F. UAV for 3D mapping applications: a review. *Applied Geomatics*, 2014; 6(1): 1 - 15. <https://doi.org/10.1007/s12518-013-0120-x>
- [7] Karkınlı AE, Janowski A, Kaderli L, Gül Hüsrevoğlu B, Hüsrevoğlu M. Robust Multi-Sensor Point Cloud Registration for Cultural Heritage Documentation: A Multi-Population Based Differential Evolution Approach. *Remote Sensing*, 2026; 18(12): 1971. <https://doi.org/10.3390/rs18121971>
- [8] Bhatnagar S, Gill L, Ghosh B. Drone Image Segmentation Using Machine and Deep Learning for Mapping Raised Bog Vegetation Communities. *Remote Sensing*, 2020; 12(16): 2602. <https://doi.org/10.3390/rs12162602>
- [9] Lendzioch T, Langhammer J, Vlček L, Minařík R. Mapping the Groundwater Level and Soil Moisture of a Montane Peat Bog Using UAV Monitoring and Machine Learning. *Remote Sensing*, 2021; 13(5): 907. <https://doi.org/10.3390/rs13050907>
- [10] Lu T, Wan L, Qi S, Gao M. Land Cover Classification of UAV Remote Sensing Based on Transformer-CNN Hybrid Architecture. *Sensors*, 2023; 23(11): 5288. <https://doi.org/10.3390/s23115288>
- [11] Amarasingam N, Vanegas F, Hele M, Warfield A, Gonzalez F. Integrating Artificial Intelligence and UAV-Acquired Multispectral Imagery for the Mapping of Invasive Plant Species in Complex Natural Environments. *Remote Sensing*, 2024; 16(9): 1582. <https://doi.org/10.3390/rs16091582>
- [12] Szostak R, Pietroni M, Wachniew P, Zimnoch M, Cwiakala P. Estimation of Small-Stream Water Surface Elevation Using UAV Photogrammetry and Deep Learning. *Remote Sensing*, 2024; 16(8): 1458. <https://doi.org/10.3390/rs16081458>
- [13] Kesikoglu MH, Ozmen HB. Impact of preprocessing in UAV images on road surface damage detection. *Research on Engineering Structures and Materials*, 2026; 12(4): 2791 - 2810. <https://doi.org/10.17515/resm2026-1355st1020rs>
- [14] Lyu Y, Vosselman G, Xia GS, Yilmaz A, Yang MY. UAVid: A semantic segmentation dataset for UAV imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2020; 165: 108 - 119. <https://doi.org/10.1016/j.isprsjprs.2020.05.009>
- [15] Long J, Shelhamer E, Darrell T. Fully convolutional networks for semantic segmentation. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015: 3431 - 3440. <https://doi.org/10.1109/CVPR.2015.7298965>
- [16] Ronneberger O, Fischer P, Brox T. U-Net: Convolutional Networks for Biomedical Image Segmentation. *Medical Image Computing and Computer-Assisted Intervention (MICCAI 2015), Lecture Notes in Computer Science*, 2015; 9351: 234 - 241. https://doi.org/10.1007/978-3-319-24574-4_28
- [17] Chen LC, Zhu Y, Papandreou G, Schroff F, Adam H. Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation. *Computer Vision (ECCV 2018), Lecture Notes in Computer Science*, 2018; 11211: 833 - 851. https://doi.org/10.1007/978-3-030-01234-2_49
- [18] Xie E, Wang W, Yu Z, Anandkumar A, Alvarez JM, Luo P. SegFormer: Simple and efficient design for semantic segmentation with transformers. *Advances in Neural Information Processing Systems*, 2021; 34: 12077 - 12090.
- [19] Yuan X, Shi J, Gu L. A review of deep learning methods for semantic segmentation of remote sensing imagery. *Expert Systems with Applications*, 2021; 169: 114417. <https://doi.org/10.1016/j.eswa.2020.114417>
- [20] Wang L, Li R, Zhang C, Fang S, Duan C, Meng X, Atkinson PM. UNetFormer: A UNet-like transformer for efficient semantic segmentation of remote sensing urban scene imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022; 190: 196 - 214. <https://doi.org/10.1016/j.isprsjprs.2022.06.008>
- [21] Kesikoglu MH. Enhancing Snow Detection through Deep Learning: Evaluating CNN Performance Against Machine Learning and Unsupervised Classification Methods. *Water Resources Management*, 2025; 39(12): 6053 - 6071. <https://doi.org/10.1007/s11269-025-04240-4>
- [22] Tobler WR. A Computer Movie Simulating Urban Growth in the Detroit Region. *Economic Geography*, 1970; 46: 234 - 240. <https://doi.org/10.2307/143141>
- [23] Roberts DR, Bahn V, Ciuti S, Boyce MS, Elith J, Guillera-Aroita G, Hauenstein S, Lahoz-Monfort JJ, Schröder B, Thuiller W, Warton DI, Wintle BA, Hartig F, Dormann CF. Cross-validation strategies for data

- p>with temporal, spatial, hierarchical, or phylogenetic structure.
- Ecography*
- , 2017; 40(8): 913 - 929.
- <https://doi.org/10.1111/ecog.02881>
- [24] Meyer H, Reudenbach C, Hengl T, Katurji M, Nauss T. Improving performance of spatio-temporal machine learning models using forward feature selection and target-oriented validation. *Environmental Modelling & Software*, 2018; 101: 1 - 9. <https://doi.org/10.1016/j.envsoft.2017.12.001>
 - [25] Ploton P, Mortier F, Réjou-Méchain M, Barbier N, Picard N, Rossi V, Dormann C, Cornu G, Viennois G, Bayol N, Lyapustin A, Gourlet-Fleury S, Péliissier R. Spatial validation reveals poor predictive performance of large-scale ecological mapping models. *Nature Communications*, 2020; 11(1): 4540. <https://doi.org/10.1038/s41467-020-18321-y>
 - [26] Karasiak N, Dejoux JF, Monteil C, Sheeren D. Spatial dependence between training and test sets: another pitfall of classification accuracy assessment in remote sensing. *Machine Learning*, 2022; 111(7): 2715 - 2740. <https://doi.org/10.1007/s10994-021-05972-1>
 - [27] Kattenborn T, Schiefer F, Frey J, Feilhauer H, Mahecha MD, Dormann CF. Spatially autocorrelated training and validation samples inflate performance assessment of convolutional neural networks. *ISPRS Open Journal of Photogrammetry and Remote Sensing*, 2022; 5: 100018. <https://doi.org/10.1016/j.ophoto.2022.100018>
 - [28] Valavi R, Elith J, Lahoz-Monfort JJ, Guillera-Arroita G. blockCV: An R package for generating spatially or environmentally separated folds for k-fold cross-validation of species distribution models. *Methods in Ecology and Evolution*, 2019; 10(2): 225 - 232. <https://doi.org/10.1111/2041-210X.13107>
 - [29] Lyons MB, Keith DA, Phinn SR, Mason TJ, Elith J. A comparison of resampling methods for remote sensing classification and accuracy assessment. *Remote Sensing of Environment*, 2018; 208: 145 - 153. <https://doi.org/10.1016/j.rse.2018.02.026>
 - [30] Taşkanat T. Enhancing Artificial Neural Network Performance for Wildfire Susceptibility Mapping Using Bernstein-Levy and Multi-Population Differential Evolution Algorithms. *Transactions in GIS*, 2026; 30(4): e70318. <https://doi.org/10.1111/tgis.70318>
 - [31] Wadoux AMJC, Heuvelink GBM, de Bruin S, Brus DJ. Spatial cross-validation is not the right way to evaluate map accuracy. *Ecological Modelling*, 2021; 457: 109692. <https://doi.org/10.1016/j.ecolmodel.2021.109692>
 - [32] Stehman SV, Foody GM. Key issues in rigorous accuracy assessment of land cover products. *Remote Sensing of Environment*, 2019; 231: 111199. <https://doi.org/10.1016/j.rse.2019.05.018>
 - [33] He K, Zhang X, Ren S, Sun J. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016: 770 - 778. <https://doi.org/10.1109/CVPR.2016.90>
 - [34] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 2015; 115(3): 211 - 252. <https://doi.org/10.1007/s11263-015-0816-y>
 - [35] Milletari F, Navab N, Ahmadi SA. V-Net: Fully Convolutional Neural Networks for Volumetric Medical Image Segmentation. 2016 Fourth International Conference on 3D Vision (3DV), 2016: 565 - 571. <https://doi.org/10.1109/3DV.2016.79>
 - [36] Eigen D, Fergus R. Predicting Depth, Surface Normals and Semantic Labels with a Common Multi-scale Convolutional Architecture. 2015 IEEE International Conference on Computer Vision (ICCV), 2015: 2650 - 2658. <https://doi.org/10.1109/ICCV.2015.304>
 - [37] Loshchilov I, Hutter F. Decoupled weight decay regularization. *International Conference on Learning Representations (ICLR)*, 2019.
 - [38] Breiman L. Random Forests. *Machine Learning*, 2001; 45(1): 5 - 32. <https://doi.org/10.1023/A:1010933404324>
 - [39] Foody GM. Status of land cover classification accuracy assessment. *Remote Sensing of Environment*, 2002; 80(1): 185 - 201. [https://doi.org/10.1016/S0034-4257\(01\)00295-4](https://doi.org/10.1016/S0034-4257(01)00295-4)
 - [40] Everingham M, Van Gool L, Williams CKI, Winn J, Zisserman A. The Pascal Visual Object Classes (VOC) Challenge. *International Journal of Computer Vision*, 2010; 88(2): 303 - 338. <https://doi.org/10.1007/s11263-009-0275-4>
 - [41] Wilcoxon F. Individual Comparisons by Ranking Methods. *Biometrics Bulletin*, 1945; 1(6): 80 - 83. <https://doi.org/10.2307/3001968>
 - [42] Holm S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 1979; 6(2): 65 - 70.
 - [43] Nigam I, Huang C, Ramanan D. Ensemble Knowledge Transfer for Semantic Segmentation. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV), 2018: 1499 - 1508. <https://doi.org/10.1109/WACV.2018.00168>
 - [44] Karkınlı AE. Multi-scale remote sensing of desertification trends and climate-vegetation interactions in the Konya basin, Türkiye (2000-2025). *Niğde Ömer Halisdemir University Journal of Engineering Sciences*, 2025; 14(4): 1690 - 1699. <https://doi.org/10.28948/ngumuh.1791557>

- [45] Taşkanat T, Karkınlı AE, Kesikoğlu A, Beşdok E. Evolutionary optimization of neural network for landslide susceptibility assessment. *Stochastic Environmental Research and Risk Assessment*, 2025; 39(12): 6127 - 6153. <https://doi.org/10.1007/s00477-025-03118-6>
- [46] Xu C, Wang J, Yang W, Yu H, Yu L, Xia GS. Detecting tiny objects in aerial images: A normalized Wasserstein distance and a new benchmark. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2022; 190: 79 - 93. <https://doi.org/10.1016/j.isprsjprs.2022.06.002>
- [47] Wu X, Li W, Hong D, Tao R, Du Q. Deep Learning for Unmanned Aerial Vehicle-Based Object Detection and Tracking: A survey. *IEEE Geoscience and Remote Sensing Magazine*, 2022; 10(1): 91 - 124. <https://doi.org/10.1109/MGRS.2021.3115137>
- [48] Meyer H, Pebesma E. Predicting into unknown space? Estimating the area of applicability of spatial prediction models. *Methods in Ecology and Evolution*, 2021; 12(9): 1620 - 1633. <https://doi.org/10.1111/2041-210X.13650>