HSI-ROAD RELABELED: SURFACE-AWARE ROAD-SCENE SEGMENTATION

Imad Ali Shah [1,], Imran Mehmood [1], Enda Ward [2], Martin Glavin [1], Edward Jones [1] and Brian Deegan [1]*

[1] School of Engineering and Ryan Institute, University of Galway, Ireland
[2] Valeo Vision Systems, Tuam, Ireland

ABSTRACT

The HSI-Road dataset provides paired RGB and 25-channel NIR (680–960 nm) images with binary masks but no surface-level labels. This paper introduces a manually labeled six-class taxonomy: Background, Asphalt, Concrete, Dirt, Water, and Grass, and an RGB-to-NIR registration pipeline with corresponding annotations. Six semantic-segmentation models are evaluated under four input configurations: original-resolution RGB, registered low-resolution RGB (RGB_{reg}), NIR, and channel-stacked RGB_{reg}–NIR ($RGBN_{stk}$). The comparison quantifies the effect of spatial-resolution reduction on RGB, along with evaluation of NIR and $RGBN_{stk}$, with results reported using per-class and mean IoU and F1 scores. The original-resolution RGB achieves the highest overall performance but contains 12× more pixels and incurs a 15.5–20.6% latency penalty compared to the reduced-resolution inputs. At the common 192×384 resolution, $RGBN_{stk}$ outperforms NIR for all six models and RGB_{reg} for five of six, with the most consistent gains for the Water class. These results highlight the importance of spatial resolution while showing that NIR provides complementary information to RGB.

***Index Terms*—** HSI-Road, Multimodal Segmentation, NIR, Surface-oriented Road Annotations, UPerNet

1. INTRODUCTION

Road-surface understanding is an important aspect of autonomous driving (AD), especially for traction control and road-condition monitoring. The public hyperspectral imaging (HSI) driving datasets [1] generally represent the road as a single semantic class rather than distinguishing road-surface materials. This supports free-space detection but not surface-aware tasks such as distinguishing paved from unpaved terrain or detecting standing water and drivable grassy patches. Spectral imaging such as near-infrared (NIR) may complement RGB due to different materials' reflectance across visible and NIR wavelengths, helping distinguish surfaces that may appear similar in RGB. However, their robustness requires empirical evaluation for each class and downstream architecture.

HSI-Road [2] is one of the largest public HSI datasets for AD [3] and contains various drivable surfaces (such as asphalt, dirt, etc), but its binary masks do not support material-level segmentation. In addition, HSI-Road has not previously been extended with multiclass surface-oriented annotations. This gap motivates the current work, with its main contributions as follows:

1. Six-class, surface-oriented HSI-Road annotations, along with an RGB-to-NIR registration pipeline.
2. Four-input configuration-based evaluation: (i) RGB_{ori}: Original dataset resolution RGB, (ii) RGB_{reg}: Registered low-resolution RGB, (iii) NIR; and (iv) $RGBN_{stk}$: Channel-stacked RGB_{reg}–NIR.
3. Multilabel-stratified train/validation/test split that approximately preserves image-level class prevalence.

2. BACKGROUND AND MOTIVATION

2.1. RGB and NIR for Road Scene Understanding

Remote sensing widely combines NIR and RGB modalities to discriminate materials [4] such as vegetation, moisture, and mineral surfaces, leveraging their distinct spectral characteristics. However, its exploration in the AD domain remains limited [1]. Prior works using available HSI datasets [3] for road scenes generally target drivable areas (e.g., roads) versus other broader multiclass urban-object segmentation. Moreover, explicit matched-resolution comparisons between RGB_{reg} and NIR for surface-oriented multiclass segmentation remain underexplored. This gap makes it difficult to separate pure multimodal gains from performance degradation caused by registration artifacts and spatial-resolution mismatch. Recent studies demonstrate that spatial misregistration between modalities can severely degrade detection performance [5]. Consequently, explicit RGB–NIR alignment is important for pixel-level fusion approaches.

2.2. HSI-Road Dataset and HSI Segmentation in AD

HSI-Road contains 3799 paired RGB (704×1280) and 25-channel NIR (192×384) images with pixel-level binary masks: Background (Bkg.) and Road. It covers asphalt, concrete (Conc.), and dirt track, with over 60% images on rural

This work was funded, in part, by Taighde Éireann–Research Ireland grants 13/RC/2094_P2 and 18/SP/5942, and Valeo Vision Systems.
Annotations, data splits, and registration metadata are made available at `https://github.com/imadalishah/HSI_Road_relabeled`.
Correspondence: i.shah2@universityofgalway.ie

roads [2]. Despite this diversity, the binary labelling scheme restricts the dataset to drivable-area segmentation, motivating extension to road surface-oriented annotations.

Unlike extensive exploration of deep learning (DL) in other domains [4], DL in HSI for AD remains relatively limited [1]. Previous studies have considered pixel-level ANN-based classification and baseline semantic segmentation models (SSMs) [3], including UNet [6], DeepLabV3+ [7], and SegFormer [8]. These efforts demonstrate increasing interest in HSI-based road-scene semantic understanding.

Table 1: Pixel-level class distribution comparison between the original binary and the relabeled multi-class annotations.

Label	Class	NIR (192×384)		RGB (704×1280)	
		Pixels	%	Pixels	%
Original Binary Class Distribution					
0	Bkg.	204,185,446	72.90	2,513,487,955	73.42
1	Road	75,907,226	27.10	909,866,925	26.58
Relabeled Multi-Class Distribution					
0	Bkg. [1]	201,076,168	71.79	2,474,507,082	72.28
1	Asphalt	27,689,446	9.89	331,948,381	9.70
2	Conc. [2]	15,547,453	5.55	188,276,659	5.50
3	Dirt [3]	33,587,271	11.99	402,433,271	11.76
4	Water [4]	794,868	0.28	9,553,815	0.28
5	Grass [5]	1,397,466	0.50	16,635,672	0.49

[1] Include all pixels not assigned to the other five target surface classes. [2] Concrete. [3] Unpaved dirt, soil, and sand surfaces. [4] Standing water and roadside drains/ditches. [5] Roadside grassy regions considered drivable under the adopted annotation taxonomy.

3. ANNOTATION AND MODALITY ALIGNMENT

3.1. Six-Class Surface-Oriented Taxonomy

The relabeled annotation contains six mutually exclusive pixel classes: Bkg., Asphalt, Conc., Dirt, Water, and Grass. Table 1 compares the original binary and the relabeled multiclass annotations, and reports the corresponding pixel distributions. The six-class annotations were created manually using RGB_{ori}, which provides finer spatial detail than NIR, with a custom segmentation tool. The relabeled annotations are not a strict subdivision of the original Road class, and include redrawn surface boundaries under the new annotation taxonomy. Consequently, some pixels originally labeled Bkg. are assigned to surface classes, such as Water and Grass. The relabeling, therefore, modifies both semantic granularity and the local spatial extent of the annotated surface region.

3.2. Spatial Registration: RGB-to-NIR Modality

RGB was co-registered to the NIR resolution. A pseudo-NIR grayscale was formed by averaging three channels, initially using 0, 14, and 23 [1]. However, if post-registration Normalized Cross-Correlation (NCC) was < 0.70, remaining triplets among the $\binom{25}{3} - 1 = 2299$ combinations were tested until NCC≥ 0.70. If no triplet reached this threshold, the highest-NCC based triplet was retained [2].

[1] Combination selected empirically after testing 100 random images.
[2] Metadata in the codebase includes image-wise best triplets.

Contrast-enhanced grayscales were matched using Oriented FAST and Rotated BRIEF (ORB) with Scale-Invariant Feature Transform (SIFT) fallback, and Random Sample Consensus (RANSAC) estimated the initial affine transform under conservative geometric constraints. Enhanced Correlation Coefficient (ECC) refinement using Euclidean and translation models estimated a residual transform that was composed with the feature-based transform. Across all mapped RGB–NIR pairs, the mean $\pm$ SD NCC was 0.73 ± 0.12. The mean inlier reprojection residual was 1.22 ± 0.15 px on the NIR grid. Example alignments are shown in Fig. 1. The same mapping was applied to semantic masks using nearest-neighbor interpolation to preserve categorical labels.

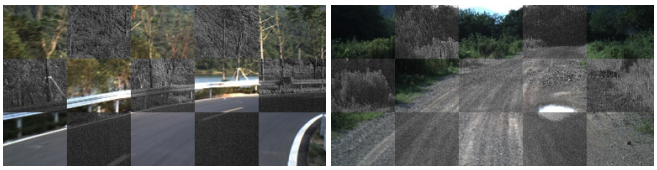


Fig. 1: Checkerboard visualization of spatial alignment using alternating RGB and NIR image patches.

Table 2: Image-level class distribution across dataset splits.

Class	Train ($N = 2659 : 70\%$)		Val ($N = 570 : 15\%$)		Test ($N = 570 : 15\%$)	
	Count	%	Count	%	Count	%
Bkg.	2659	100	570	100	570	100
Asphalt	816	30.69	175	30.70	175	30.70
Conc.	770	28.96	165	28.95	165	28.95
Dirt	1578	59.35	321	56.32	323	56.67
Water	727	27.34	156	27.37	156	27.37
Grass	484	18.20	103	18.07	104	18.25

Table 3: Training hyperparameters and input configurations.

Hyperparameter	Value / Range
Optimization & Training	
Epochs / Early Stop	Max 300 / 50 epochs on validation mIoU
Batch Size / Seeds	16 / 3 (42, 43, 44)
Optimizer / Rate	AdamW / 1×10^{-4}
Loss Function	Avg of Cross-Entropy and Dice Loss
Data split	Stratified 70/15/15 (see in Table 2)
Precision	Automatic mixed precision
SSMs Input (C, H, W)	
RGB_{ori}	(3, 704, 1280)
RGB_{reg} and NIR	(3, 192, 384) and (25, 192, 384)
RGBN_{stk}	(28, 192, 384)
Horizontal/Vertical Flip	$p = 0.5$ (at dataloader)
Hardware & Environment	
GPU	NVIDIA RTX PRO 6000 Blackwell Max-Q Edition
CPU	Dual Intel Xeon Gold 6252 (48C/96T, 2.10 GHz)
OS/workers	Debian 12 and 4 Data Loader Workers

4. EXPERIMENTAL SETUP AND RESULTS

4.1. Segmentation Backbones

Six SSMs are evaluated on the relabeled annotations of the HSI-Road dataset: UNet, UNet-CBAM [3, 9], DeepLabV3+, SegFormer, and two UPerNet [10] variants using EfficientNet-B0 and MiT-B0 encoders. UNet, UNet-CBAM, and DeepLabV3+

use EfficientNet-B0, whereas SegFormer uses MiT-B0. To evaluate the effect of encoder choice while holding the decoder fixed, UPerNet is evaluated with EfficientNet-B0 ($\text{UPerNet}_{\text{EN-B0}}$) as the CNN and MiT-B0 ($\text{UPerNet}_{\text{MiT-B0}}$) as the transformer encoder. For NIR and RGBN_{stk}, the first layer is adapted to accept 25 and 28 channels, respectively.

4.2. Data Split, Model Training and Evaluation

The 3799 images are divided into train/validation/test subsets containing 2659/570/570 images, respectively, as shown in Table 2, using multilabel stratification based on image-level class presence. Since the image-wise split is not provided in the original dataset, the codebase of this work includes it.

All SSMs use the training settings reported in Table 3 and are trained three times from scratch using different seeds. All six SSMs are evaluated using four input configurations: (1) RGB_{ori}; (2) RGB_{reg}; (3) NIR; and (4) RGBN_{stk}. The RGB_{ori}–to–RGB_{reg} comparison reflects the combined effects of spatial resizing, registration, interpolation, residual misregistration, and transformation of the corresponding relabeled masks. RGB_{reg}, NIR, and RGBN_{stk} use the same 192×384 spatial resolution and therefore provide matched resolution-based input modality comparisons. In addition to evaluation on relabeled masks, the experiments also include original binary labels on RGB_{ori}, NIR, and RGBN_{stk} input modalities.

Table 4: Modality-wise evaluation on the binary taxonomy. mIoU/mF1: dataset-pooled* macro mean±SD across seeds.

Modality	SSM	mIoU	mF1	IoU		F1	
				Bkg.	Road	Bkg.	Road
RGB_{ori}	UNet	96.86±0.04	98.40±0.02	98.32±0.02	95.41±0.05	99.15±0.01	97.65±0.03
	UNet-CBAM	96.77±0.06	98.36±0.03	98.28±0.03	95.27±0.09	99.13±0.02	97.58±0.05
	DeepLabV3+	**96.94±0.02**	**98.44±0.01**	**98.37±0.02**	**95.51±0.03**	**99.18±0.01**	**97.70±0.02**
	SegFormer	96.76±0.03	98.35±0.01	98.26±0.01	95.25±0.04	99.12±0.01	97.57±0.02
	$\text{UPerNet}_{\text{EN-B0}}$	96.89±0.05	98.42±0.02	98.34±0.03	95.44±0.07	99.16±0.01	97.67±0.03
	$\text{UPerNet}_{\text{MiT-B0}}$	96.90±0.05	98.42±0.03	98.34±0.03	95.45±0.07	99.16±0.01	97.67±0.04
NIR	UNet	95.84±0.11	97.87±0.06	97.71±0.06	93.97±0.15	98.84±0.03	96.89±0.08
	UNet-CBAM	95.81±0.12	97.85±0.06	97.69±0.07	93.93±0.17	98.83±0.04	96.87±0.09
	DeepLabV3+	95.87±0.14	97.88±0.07	97.73±0.07	94.01±0.20	98.85±0.04	96.91±0.11
	SegFormer	95.89±0.03	97.89±0.02	97.74±0.02	94.04±0.04	98.86±0.01	96.93±0.02
	$\text{UPerNet}_{\text{EN-B0}}$	95.97±0.02	97.93±0.01	97.79±0.01	94.15±0.03	98.88±0.00	96.99±0.01
	$\text{UPerNet}_{\text{MiT-B0}}$	**96.02±0.02**	**97.96±0.01**	**97.82±0.01**	**94.22±0.03**	**98.90±0.01**	**97.03±0.02**
RGBN_{stk}	UNet	96.09±0.02	98.00±0.01	97.85±0.01	94.32±0.04	98.92±0.00	97.08±0.02
	UNet-CBAM	95.95±0.05	97.92±0.03	97.78±0.03	94.12±0.07	98.88±0.02	96.97±0.04
	DeepLabV3+	96.01±0.09	97.95±0.05	97.80±0.05	94.21±0.13	98.89±0.03	97.02±0.07
	SegFormer	96.04±0.05	97.97±0.03	97.82±0.03	94.25±0.07	98.90±0.02	97.04±0.04
	$\text{UPerNet}_{\text{EN-B0}}$	96.19±0.04	98.05±0.02	97.91±0.02	94.46±0.05	98.94±0.01	97.15±0.03
	$\text{UPerNet}_{\text{MiT-B0}}$	**96.21±0.06**	**98.06±0.03**	**97.92±0.03**	**94.49±0.08**	**98.95±0.02**	**97.17±0.04**

* For each seed, class-wise metrics are pooled over all pixels in the test set.

4.3. Overall Performance

Tables 4 and 5 report performance on the original binary and relabeled six-class masks, respectively. Binary mask performance is consistently high and tightly clustered, with 95.81–96.94% mIoU and 97.85–98.44% mF1 across all evaluated SSM-to-modality combinations, consistent with prior studies [3]. For RGB_{ori}, DeepLabV3+ achieves the highest mIoU at 96.94%, followed by $\text{UPerNet}_{\text{MiT-B0}}$ at 96.90%. $\text{UPerNet}_{\text{MiT-B0}}$ also achieves the highest binary mIoU for NIR of 96.02% and for RGBN_{stk} of 96.21%. However, the narrow spread of 1.13 mIoU and 0.59 mF1 points indicates that the results on binary masks provide limited performance insights across the evaluated SSM-to-modality combinations.

In contrast, the six-class taxonomy yields a wider performance range, with mIoU from 73.39–82.73% and mF1 from 81.70–89.43%. The remainder of the analysis therefore focuses on the six-class taxonomy only. $\text{UPerNet}_{\text{MiT-B0}}$ achieves the highest mIoU for three input configurations (RGB_{ori}: 82.73± 0.40%, RGB_{reg}: 77.03±0.27%, and RGBN_{stk}: 79.66±0.46%), while $\text{UPerNet}_{\text{EN-B0}}$ performs best for NIR with 77.37±0.38%. Across input modalities, Bkg., Asphalt, Conc., and Dirt retain relatively high IoU, whereas Water and Grass are substantially more difficult to segment. Although the best RGBN_{stk} SSM remains 3.07 mIoU points below RGB_{ori}, the two operate at different spatial resolutions. At the matched 192×384 resolution, RGBN_{stk} exceeds RGB_{reg} by 2.63 mIoU points for $\text{UPerNet}_{\text{MiT-B0}}$, indicating that NIR complements the RGB modality and increases performance.

4.4. Effect of Spatial Mapping: RGB_{ori} vs. RGB_{reg}

Mapping RGB_{ori} (704×1280) to the registered NIR resolution (192×384) induces a near 12x difference in the number of pixels, resulting in a performance drop of RGB_{reg} for all classes across the evaluated SSMs, as found in Table 5. The reductions in mIoU points are: 4.44 for UNet, 4.01 for UNet-CBAM, 4.72 for DeepLabV3+, 4.67 for SegFormer, 5.97 for $\text{UPerNet}_{\text{EN-B0}}$, and 5.70 for $\text{UPerNet}_{\text{MiT-B0}}$, for an average drop of 4.92 points. The UPerNet variants exhibit the largest absolute reductions despite achieving the best performance.

The effect is strongly class-dependent, averaged across SSMs: Water drops IoU by 12.68 points and Grass by 8.89 points, compared with 2.95 for Conc., 2.34 for Asphalt, 2.13 for Dirt, and only 0.51 for Bkg. Water and Grass are therefore the classes most affected by the changes in spatial resolution and applied registration. While this work establishes a strong baseline across modalities, isolating the effects of spatial resolution or boundary complexity remains for future work. Nonetheless, the newly aligned RGB-NIR offers a valuable modality for future multimodal research on HSI-Road.

4.5. Matched resolution: NIR vs. RGB_{reg} vs. RGBN_{stk}

NIR vs. RGB_{reg}: At the matched 192×384 spatial resolution, NIR outperforms RGB_{reg} in two SSMs (SegFormer by 0.94 mIoU points and $\text{UPerNet}_{\text{EN-B0}}$ by 1.09), but decreases performance in four (UNet: 2.03, UNet-CBAM: 2.06, DeepLabV3+: 1.42, and $\text{UPerNet}_{\text{MiT-B0}}$: 3.61). The $\text{UPerNet}_{\text{MiT-B0}}$ on NIR also exhibits the lowest consistency at $73.42 \pm 3.11\%$ mIoU, indicating greater run-to-run variability than the other evaluated configurations.

RGBN_{stk} vs. RGB_{reg}: RGBN_{stk} achieves higher mean mIoU than RGB_{reg} for five of the six SSMs (UNet: 1.44, DeepLabV3+: 1.24, SegFormer: 3.52, $\text{UPerNet}_{\text{EN-B0}}$: 3.11, and $\text{UPerNet}_{\text{MiT-B0}}$: 2.63), whereas UNet-CBAM decreases slightly by 0.14 points. For the five SSMs, RGBN_{stk} narrows the mIoU gap between RGB_{ori} and RGB_{reg}.

Table 5: Modality-wise evaluation on the six-class taxonomy. mIoU/mF1: dataset-pooled macro mean±SD across seeds.

Modality	SSMs	mIoU	mF1	IoU						F1					
				Bkg.	Asphalt	Conc.	Dirt	Water	Grass	Bkg.	Asphalt	Conc.	Dirt	Water	Grass
RGB_{ori}	UNet	79.86±0.23	87.19±0.20	97.67±0.01	96.10±0.29	92.88±0.51	90.23±0.40	47.34±0.39	54.94±2.27	98.82±0.00	98.01±0.15	96.31±0.27	94.87±0.22	64.26±0.35	70.90±1.89
	UNet-CBAM	79.78±0.47	87.17±0.47	97.67±0.02	95.81±0.49	92.59±0.49	90.09±0.48	48.03±2.78	54.49±1.48	98.82±0.01	97.86±0.26	96.15±0.26	94.79±0.27	64.86±2.52	70.53±1.25
	DeepLabV3+	81.33±0.86	88.37±0.62	97.83±0.03	96.08±0.74	93.02±1.08	91.07±0.21	49.81±2.14	60.18±2.75	98.91±0.02	98.00±0.38	96.38±0.58	95.32±0.11	66.48±1.90	75.12±2.16
	SegFormer	79.37±0.62	86.84±0.54	97.61±0.01	95.54±0.54	91.73±0.68	90.36±0.31	46.02±2.37	54.95±1.40	98.79±0.01	97.72±0.28	95.69±0.37	94.93±0.17	63.00±2.24	70.92±1.17
	$UPerNet_{EN-B0}$	82.25±0.55	89.02±0.46	97.93±0.01	96.55±0.19	**93.79±0.34**	**91.61±0.04**	51.09±2.49	62.50±0.71	98.96±0.00	98.24±0.10	**96.79±0.18**	**95.62±0.02**	67.60±2.20	76.92±0.54
	$UPerNet_{MiT-B0}$	**82.73±0.40**	**89.43±0.33**	**97.97±0.03**	**96.73±0.18**	93.77±0.33	91.51±0.16	**53.25±1.25**	**63.18±1.80**	**98.97±0.01**	**98.34±0.10**	96.78±0.17	95.57±0.09	**69.49±1.06**	**77.42±1.36**
RGB_{reg}	UNet	75.42±1.01	83.49±0.76	97.20±0.05	93.49±1.03	89.64±1.58	88.61±0.70	35.14±2.88	48.47±3.42	98.58±0.03	96.63±0.55	94.53±0.88	93.96±0.40	51.96±3.16	65.25±3.10
	UNet-CBAM	75.77±1.59	83.60±1.56	97.16±0.11	94.40±0.52	91.02±0.82	88.51±0.64	33.84±5.94	49.67±1.87	98.56±0.06	97.12±0.27	95.30±0.45	93.91±0.36	50.37±6.78	66.36±1.67
	DeepLabV3+	76.61±0.85	84.41±0.75	97.22±0.08	**94.50±0.60**	**91.11±0.78**	**89.20±0.25**	36.23±2.60	**51.39±1.28**	98.59±0.04	**97.17±0.32**	**95.35±0.43**	**94.29±0.14**	53.16±2.82	**67.89±1.11**
	SegFormer	74.70±0.81	83.05±0.72	97.18±0.09	92.28±0.60	87.70±1.33	87.96±0.36	34.96±1.96	48.10±1.91	98.57±0.05	95.99±0.32	93.45±0.76	93.60±0.21	51.79±2.14	64.94±1.73
	$UPerNet_{EN-B0}$	76.28±1.06	84.22±1.01	97.40±0.05	94.16±0.86	90.34±0.53	88.88±0.74	38.40±4.68	48.54±4.85	98.68±0.02	96.99±0.46	94.92±0.29	94.11±0.42	55.38±4.97	65.26±4.48
	$UPerNet_{MiT-B0}$	**77.03±0.27**	**84.99±0.24**	**97.45±0.06**	93.94±0.23	90.25±1.05	88.92±0.35	**40.87±0.75**	50.74±2.29	**98.71±0.03**	96.88±0.12	94.88±0.58	94.13±0.19	**58.02±0.75**	67.30±2.00
NIR	UNet	73.39±0.86	81.70±0.92	96.66±0.14	93.05±0.41	88.35±2.33	87.36±0.58	31.82±4.23	43.12±1.56	98.30±0.07	96.40±0.22	93.80±1.32	93.25±0.33	48.17±4.84	60.24±1.53
	UNet-CBAM	73.71±1.42	82.02±1.17	96.59±0.18	92.48±1.61	89.23±1.04	87.10±1.79	32.28±2.46	44.55±2.70	98.26±0.09	96.09±0.87	94.31±0.58	93.10±1.03	48.77±2.79	61.61±2.56
	DeepLabV3+	75.19±0.42	83.31±0.32	96.72±0.09	93.22±0.50	**89.86±1.45**	88.43±0.42	34.79±0.60	48.11±3.19	98.33±0.05	96.49±0.27	**94.65±0.80**	93.86±0.24	51.61±0.66	64.92±2.90
	SegFormer	75.64±1.44	83.85±1.21	96.85±0.17	93.45±0.79	89.43±1.49	**88.49±1.02**	40.98±2.02	44.66±4.15	98.40±0.09	96.61±0.42	94.42±0.83	**93.89±0.58**	58.12±2.03	61.67±3.94
	$UPerNet_{EN-B0}$	**77.37±0.38**	**85.51±0.23**	**96.99±0.07**	**93.48±0.57**	88.94±1.99	88.47±0.50	**45.64±0.89**	**50.69±1.19**	**98.47±0.03**	**96.63±0.30**	94.14±1.12	93.88±0.28	**62.67±0.84**	**67.27±1.04**
	$UPerNet_{MiT-B0}$	73.42±3.11	82.06±2.97	96.50±0.87	91.90±0.94	86.46±2.12	85.60±4.17	36.61±11.76	43.47±2.53	98.22±0.45	95.78±0.51	92.73±1.23	92.20±2.44	52.85±13.16	60.57±2.46
$RGBN_{stk}$	UNet	76.86±0.66	84.70±0.52	97.19±0.06	94.92±0.42	91.33±0.79	89.20±0.50	40.40±2.38	48.13±1.94	98.58±0.03	97.39±0.22	95.47±0.43	94.29±0.28	57.53±2.40	64.96±1.75
	UNet-CBAM	75.63±0.98	83.57±0.87	97.12±0.09	94.43±0.95	90.85±1.14	89.07±0.13	37.86±0.34	44.45±5.22	98.54±0.05	97.13±0.50	95.20±0.62	94.22±0.07	54.93±0.36	61.42±4.94
	DeepLabV3+	77.85±0.83	85.51±0.78	97.29±0.05	95.00±0.06	**91.81±0.58**	89.78±0.17	40.65±3.51	52.60±1.37	98.63±0.03	97.44±0.03	**95.73±0.32**	94.61±0.09	57.74±3.55	68.94±1.18
	SegFormer	78.22±0.15	85.97±0.12	97.28±0.07	94.42±0.62	91.16±0.68	90.12±0.37	44.96±1.93	51.42±1.48	98.62±0.03	97.13±0.33	95.38±0.37	94.80±0.20	62.01±1.82	67.91±1.28
	$UPerNet_{EN-B0}$	79.39±0.63	87.00±0.56	97.45±0.02	94.48±0.24	91.09±0.40	90.11±0.27	48.60±2.60	**54.64±2.57**	98.71±0.01	97.16±0.13	95.34±0.22	94.80±0.15	65.38±2.38	**70.64±2.13**
	$UPerNet_{MiT-B0}$	**79.66±0.46**	**87.17±0.40**	**97.49±0.11**	**95.18±0.23**	91.51±0.51	**90.50±0.43**	**49.68±1.76**	53.62±2.76	**98.73±0.05**	**97.53±0.12**	95.57±0.28	**95.01±0.24**	**66.37±1.58**	69.78±2.33

Class-level Performance: The largest consistently positive class-level gains are observed for Water, for which $RGBN_{stk}$ exceeds RGB_{reg} by 4.02–10.20 points for all six SSMs. Asphalt and Dirt also show higher IoU for all six SSMs, but with smaller gains, while Conc. improves for five of six SSMs. Grass improves for four SSMs and decreases for UNet and UNet-CBAM, indicating a relatively less consistent benefit. Among the three matched-resolution modalities, $RGBN_{stk}$ achieves the highest performance, although the best-performing SSM–modality configuration varies by class. This result supports the potential value of complementary RGB–NIR information.

Table 6: Model complexity and forward-pass latencies.

Modality	SSMs	Params (M)	Mean Latency *		Peak GPU
			CPU (s)	GPU (ms)	Mem (MiB)
RGB_{ori}	UNet	6.25	4.78	12.50	445.65
	DeepLabV3+	4.91	4.77	10.56	337.95
	SegFormer	3.72	3.63	10.02	695.44
	$UPerNet_{MiT-B0}$	10.73	4.27	14.24	1002.60
	$UPerNet_{EN-B0}$	11.61	4.28	14.35	1003.13
RGB_{reg}	UNet	6.25	4.43	11.47	91.53
	DeepLabV3+	4.91	4.49	10.26	86.93
	SegFormer	3.72	2.95	9.22	99.89
	$UPerNet_{MiT-B0}$	10.73	3.47	10.36	166.94
	$UPerNet_{EN-B0}$	11.61	3.46	10.30	166.94
NIR	UNet	6.26	4.39	11.03	98.58
	DeepLabV3+	4.92	4.54	10.69	93.99
	SegFormer	3.75	3.02	8.69	106.21
	$UPerNet_{MiT-B0}$	10.77	3.46	10.50	173.54
	$UPerNet_{EN-B0}$	11.62	3.48	10.35	173.54
$RGBN_{stk}$	UNet	6.26	4.36	11.04	99.27
	DeepLabV3+	4.92	4.49	10.27	94.58
	SegFormer	3.75	2.96	8.75	107.63
	$UPerNet_{MiT-B0}$	10.77	3.47	10.18	176.03
	$UPerNet_{EN-B0}$	11.62	3.48	10.33	176.15

* Averaged over test set (570 images) with batch size 1, after 10 warmup iterations.

4.6. Backbone Comparison

Overall Performance: UPerNet variants achieve the highest mIoU and mF1 across input modalities: $UPerNet_{MiT-B0}$ leads for RGB_{ori}, RGB_{reg}, and $RGBN_{stk}$, whereas $UPerNet_{EN-B0}$ performs best for NIR. Among non-UPerNet SSMs, SegFormer with $RGBN_{stk}$ achieves the strongest result, reaching 78.22% mIoU and 85.97% mF1, and also exhibits the largest gain, improving over RGB_{reg} by 3.52 mIoU and 2.92 mF1 points. The gap between the two UPerNet variants, 3.95 mIoU and 3.45 mF1 points, on NIR highlights the importance of encoder choice even when the decoder is fixed.

Spatial-Mapping Sensitivity: All SSMs exhibit reduced performance when RGB_{ori} is mapped to RGB_{reg}, with decreases of 4.01–5.97 mIoU and 3.57–4.80 mF1 points. UNet-CBAM exhibits the smallest reduction (4.01 mIoU and 3.57 mF1 points). For RGB_{reg}, $UPerNet_{MiT-B0}$ combines the highest mean performance with the lowest cross-seed variability (77.03±0.27 mIoU and 84.99±0.24 mF1). At RGB_{ori}, it achieves the highest mean performance, although UNet exhibits lower cross-seed variability.

4.7. Computational Efficiency

Table 6 shows forward-pass latency and computational requirements on the evaluated hardware (Table 3) for each SSM–modality configuration. SegFormer has the fewest parameters (3.72–3.75M) and the lowest CPU and GPU latency across all input modalities, including 2.96s on CPU and 8.75ms on GPU for $RGBN_{stk}$. The UPerNet variants use 10.73–11.62M parameters, 2.87–3.12× the parameters in SegFormer, while achieving higher six-class segmentation accuracy. For $RGBN_{stk}$, the UPerNet variants require 10.18–10.33ms GPU latency, compared with 8.75ms for SegFormer. DeepLabV3+ has the lowest peak GPU-memory consumption across all inputs, whereas SegFormer has the lowest parameter count and inference latency. Moreover, $RGBN_{stk}$

reduces peak GPU memory by 72.0% (DeepLabV3+) up to 84.5% (SegFormer) compared to RGB_{ori}. The results therefore demonstrate a trade-off between segmentation accuracy, latency, parameter count, and memory requirements.

5. LIMITATIONS AND FUTURE WORK

Training from scratch. All SSMs are trained from scratch in this work to maintain a consistent comparison across input modalities. Future work should investigate HSI-specific pretrained vision and hyperspectral foundation models.

RGB–NIR stacking. RGBN_{stk} uses channel stacking: Intermediate, late, and attention-based fusion approaches may exploit cross-modal structure more effectively. Band selection and dimensionality reduction should also be investigated due to HSI's inherent adjacent-band redundancies.

Registration and resolution are confounded. The current work does not isolate the individual contributions of spatial resolution (Table 3), class distribution (Table 1), boundary complexity, or data splitting strategy (Table 2). Isolating these confounding factors remains a subject for future work.

6. CONCLUSION

This work extends the HSI-Road dataset with (i) a manually annotated six-class surface-oriented taxonomy, (ii) a pipeline for registering RGB images and corresponding relabeled masks to the NIR resolution, and (iii) a reference multilabel-stratified train/validation/test split. Experiments show that segmentation performance on the original binary masks is tightly clustered, whereas the six-class taxonomy reveals clearer differences among input modalities and SSMs.

RGB_{ori} produces the highest overall result, both at the SSM level and per class, with $\text{UPerNet}_{\text{MiT-B0}}$ reaching 82.73% mIoU and 89.43% mF1. Mapping RGB_{ori} to RGB_{reg} reduces mIoU by 4.01–5.97 points and mF1 by 3.57–4.80 points across all SSMs. However, RGB_{ori} (704×1280) has $12\times$ the pixel count of RGB_{reg} and RGBN_{stk} requires 3.57–$6.46\times$ less peak GPU memory than RGB_{ori}. RGB_{ori} should therefore be regarded as a high-resolution reference rather than a matched-resolution test of RGB against other input modalities.

At the matched 192×384 resolution, RGBN_{stk} improves over NIR for all six SSMs and over RGB_{reg} for five of six SSMs, with gains of up to 3.52 mIoU and 2.92 mF1 points over RGB_{reg}. Water exhibits the most consistent class-level improvement, whereas the effect on Grass remains SSM-dependent. Among the evaluated SSMs, the UPerNet variants achieve the highest mean segmentation performance, while SegFormer offers the lowest parameter count and measured inference latency, alongside the largest RGBN_{stk} gain.

These results suggest three practical considerations for surface-oriented road-scene segmentation: (i) modality comparisons should be made at matched resolution, since spatial resolution has a substantial effect on performance; (ii) once resolution is matched, complementary RGB–NIR information can yield a measurable but SSM- and class-dependent benefit; and (iii) SSM–Encoder choice should be driven by deployment constraints as much as by segmentation accuracy.

7. REFERENCES

[1] I. A. Shah, J. Li, R. George, T. Brophy, E. Ward, M. Glavin, E. Jones, and B. Deegan, "Hyperspectral sensors and autonomous driving: Technologies, limitations, and opportunities," *IEEE Open Journal of Vehicular Technology*, vol. 7, pp. 124–143, 2026.

[2] J. Lu, H. Liu, Y. Yao, S. Tao, Z. Tang, and J. Lu, "Hsi road: a hyper spectral image dataset for road segmentation," in *2020 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2020, pp. 1–6.

[3] I. A. Shah, J. Li, M. Glavin, E. Jones, E. Ward, and B. Deegan, "Hyperspectral imaging-based perception in autonomous driving scenarios: Benchmarking baseline semantic segmentation models," in *2024 14th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, 2024, pp. 1–5.

[4] D. Valme, A. Rassõlkin, and D. C. Liyanage, "From ADAS to material-informed inspection: Review of hyperspectral imaging applications on mobile ground robots," *Sensors*, vol. 25, no. 8, p. 2346, 2025.

[5] M. Usama, I. A. Shah, R. George, E. Ward, F. Collins, W. O'Grady, M. Glavin, B. Deegan, and E. Jones, "Impact of inter-modality misalignment on multispectral object detection," *IEEE Open Journal of Vehicular Technology*, vol. 7, pp. 1110–1123, 2026.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[7] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 4, pp. 834–848, 2017.

[8] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[9] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 3–19.

[10] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *European Conference on Computer Vision*. Springer, 2018.