

Article

Evaluating Architecture Scalability and Transfer Learning in Urban Scene Segmentation Using Explainable AI

Tanmay Sunil Hatkar, Abhinav Pandey and Saad B. Ahmed *

Department of Computer Science, Lakehead University, Thunder Bay, ON P7B 5E1, Canada

* Correspondence: sbinahm@lakeheadu.ca

Abstract

Semantic segmentation plays a pivotal role in autonomous driving, enabling pixel-level understanding of road scenes. Although transformer-based models such as SegFormer have shown exceptional performance on large datasets, their generalization to smaller and geographically diverse datasets remains underexplored. In this work, we analyze the scalability and transferability of SegFormer variants (B3, B4, B5) using CamVid as the base dataset. We perform cross-dataset transfer learning to KITTI and IDD, evaluate class-level performance, and explore explainable AI via confidence heatmaps. Our findings show that SegFormer-B5 achieves the highest accuracy (82.4% mIoU) on CamVid, while transfer learning from CamVid improves mIoU on KITTI by 2.57% and enhances class-specific predictions in IDD by over 70%. These results highlight the practical potential of SegFormer in real-world segmentation systems and the interpretability benefits of confidence-based visual analysis.

Keywords: semantic segmentation; transformer; transfer learning; urban scene understanding; explainable AI; confidence heatmaps; SegFormer

1. Introduction

Semantic segmentation provides fine-grained pixel-level understanding of visual scenes, which is critical for safe and robust perception in autonomous vehicles. Traditionally, convolutional neural networks (CNNs) have dominated this field, but transformer-based architectures have recently gained traction due to their ability to model long-range dependencies and global context effectively.

SegFormer, a transformer-based architecture, introduces a hierarchical encoder and lightweight decoder design that eliminates the need for positional embeddings and heavy up-sampling. Although extensively evaluated on datasets like Cityscapes, its behavior on smaller and geographically diverse datasets such as CamVid, KITTI, and IDD is less understood.

This paper addresses two major gaps in current research:

- Understanding how different SegFormer variants scale on a smaller dataset like CamVid in terms of performance and computational cost.
- Investigating the effectiveness of cross-dataset transfer learning from CamVid to KITTI (Germany) and IDD (India), each representing structured and unstructured driving environments, respectively.

Additionally, we introduce explainability into the evaluation pipeline using confidence heatmaps, allowing us to visually interpret the uncertainty of the model and the quality of a decision at the pixel level.



Academic Editor: Shuping He

Received: 11 January 2026

Revised: 12 February 2026

Accepted: 27 February 2026

Published: 1 March 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

Unlike prior studies that typically focus on architectural innovation, large-scale benchmarks, or synthetic-to-real adaptation, this work emphasizes a unified evaluation of architecture scaling, cross-dataset transfer, and explainability under limited-data conditions. In particular, the use of CamVid as a source domain reflects a practical scenario where only a small, well-annotated dataset is available for pretraining. This setup allows us to investigate how transformer-based segmentation models behave when transferred across geographically diverse environments without relying on large-scale source datasets.

This paper addresses the identified gaps as follows:

- A systematic evaluation of SegFormer variants (B3, B4, B5) on CamVid to study architecture scaling effects.
- Implementation of cross-dataset transfer learning from CamVid to KITTI and IDD, including custom class mapping strategies.
- Introduction of confidence heatmaps to visualize model prediction certainty and aid explainability in safety-critical contexts.

Rather than proposing a new architecture, this paper contributes a structured empirical analysis that combines scaling behavior, real-to-real cross-dataset transfer, and interpretability within a single experimental framework. This integrated perspective distinguishes our study from prior works that typically examine these aspects independently.

The remainder of this paper is organized as follows: Section 2 reviews prior work; Section 3 details our methodology; Section 4 describes the datasets and class mappings; Section 5 outlines evaluation metrics and experimental results; and Section 6 concludes the paper with insights and future directions.

2. Related Work

Semantic segmentation has progressed substantially with the advent of deep learning, evolving from convolution-based architectures to transformer-driven designs. Early approaches such as Fully Convolutional Networks (FCNs) and SegNet [1] introduced encoder–decoder frameworks that extracted spatial features using convolutional layers. Later models like the DenseNet-based Tiramisu [2] improved training stability via dense skip connections and deep supervision. These CNN-based models were effective for structured scene parsing but often struggled with modeling long-range dependencies and global context.

To address real-time constraints in autonomous systems, efficient architectures such as BiSeNet [3] adopted dual path designs to balance spatial precision and receptive field. PIDNet [4] leveraged principles from control theory to better handle multi-scale information, while RTFormer [5] demonstrated that transformer-based designs could match CNN efficiency in real-time settings. These approaches prioritized inference speed, but often at the expense of segmentation accuracy in complex scenes.

Transformer-based segmentation models marked a turning point, with SegFormer [6] emerging as a notable breakthrough. Its hierarchical transformer encoder, overlapped patch embeddings, and Mix-FFN layers enabled strong global context modeling while preserving local structure, outperforming convolutional models on Cityscapes with fewer parameters. Further innovations like Skip-SegFormer [7] and CFF-SegFormer [8] extended this architecture with improved multi-scale feature fusion and decoder efficiency.

Parallel to architectural advancements, transfer learning has become essential for scenarios where labeled data is scarce. DAFormer [9] showed that domain-adaptive segmentation using transformers could generalize well across synthetic and real domains. However, most prior works focus on large-scale benchmarks or synthetic-to-real adaptation. Cross-dataset generalization between small, real-world datasets—such as CamVid, KITTI, and IDD—remains underexplored. This is particularly relevant for developing perception

systems intended for deployment across diverse geographic regions. Our study addresses this gap by systematically evaluating SegFormer's transferability from CamVid to both KITTI and IDD.

Interpretability is increasingly recognized as a key aspect of trustworthy segmentation. Bayesian SegNet [10] pioneered uncertainty modeling using Monte Carlo dropout. More recent approaches explore class activation maps (CAMs), attention visualizations, and confidence heatmaps to make dense predictions more interpretable. While explainable AI (XAI) has advanced for classification tasks, its integration with transformer-based segmentation remains limited. Recent transformer-specific interpretability techniques such as attention rollout and token-based saliency mapping provide complementary perspectives, but our focus remains on confidence-driven and gradient-based visualization within the SegFormer framework. Our work contributes to this area by introducing confidence heatmaps as a diagnostic tool to expose model uncertainty and failure points in urban scene segmentation.

Despite these advances, many studies evaluate models either on large-scale benchmarks or within a single methodological dimension, leaving open questions regarding scalability on smaller datasets, real-to-real transfer across geographic domains, and the interpretability of transformer-based segmentation models.

In summary, prior research has laid the foundation for efficient, accurate, and adaptive segmentation. Yet, the combined challenges of scaling transformer models, transferring them across real-world datasets, and interpreting their predictions remain open. We address this intersection by evaluating SegFormer's performance under scaling, cross-dataset transfer, and explainability constraints, bringing a unified perspective to transformer-based semantic segmentation.

3. Datasets and Cross-Dataset Mapping

To evaluate both architectural scalability and cross-domain transferability, we used three publicly available urban driving datasets: CamVid, KITTI, and IDD. These datasets represent different geographical locations, label taxonomies, and annotation densities, making them suitable benchmarks for our study.

The Cambridge-driving Labeled Video Database (CamVid) [11] contains 701 densely annotated frames (960×720 resolution) extracted from video sequences captured in the UK. It includes 32 semantic classes covering the road, buildings, pedestrians, vehicles, sky, and street furniture. We use the standard split: 367 training, 101 validation, and 233 testing images. Due to its clean annotation and balanced scene composition, CamVid serves as the source domain for model scaling and transfer learning.

The KITTI Semantic Segmentation Benchmark [12] offers 200 high-resolution images (1242×375) from urban driving scenes in Germany. It follows a 19-class taxonomy derived from Cityscapes, focusing on structured environments with well-defined object boundaries. Due to its relatively small sample size and partial label overlap with CamVid, KITTI is selected as a target domain for cross-dataset transfer.

The Indian Driving Dataset (IDD) captures diverse and unstructured driving scenes in India. It contains 10,004 images annotated with 27 semantic classes, including region-specific categories such as *autorickshaw*, *guard rail*, and *billboard*. For consistency and computational feasibility, we use a subset comprising 1761 training and 350 validation samples. IDD introduces significant visual domain shifts, including varied lighting, occlusions, and unconventional vehicle types, making it a challenging yet valuable target domain.

To facilitate effective transfer learning from CamVid to target datasets, we establish a structured class mapping protocol. Since each dataset follows its own label taxonomy, alignment is necessary to preserve semantic consistency and ensure correct feature adapta-

tion. Our mapping strategy categorizes class relationships into three types, as summarized in Table 1.

Table 1. Cross-dataset class mapping strategy.

Mapping Type	Description	Examples
Direct Mappings	Classes with equivalent semantics and visual representation across datasets	Road → Road, Building → Building, Sky → Sky
Semantic Mappings	Classes where labels differ but visual category is similar	Bicyclist (CamVid) → Rider (KITTI, IDD), Pedestrian → Person, Vegetation and Tree → Vegetation
Unique/Novel Classes	Classes present only in target dataset with no source equivalent	KITTI-Specific: Train, Motorcycle, Terrain; IDD-Specific: Autorickshaw, Billboard, Animal

To handle Novel Classes, we reinitialize the decoder weights while preserving the CamVid-pretrained encoder. This allows the network to reuse generalized features and adapt them to new class boundaries and appearance patterns during fine-tuning. This strategy ensures semantic consistency while enabling flexible cross-dataset transfer in geographically and structurally diverse urban scenes.

4. Methodology

Our proposed methodology investigates the scalability, transferability, and interpretability of transformer-based segmentation using SegFormer as seen in the Figure 1. The experimental design is divided into three major stages: (1) model scaling experiments using CamVid; (2) cross-dataset transfer learning to KITTI and IDD; and (3) application of confidence-based interpretability techniques to evaluate model reliability.

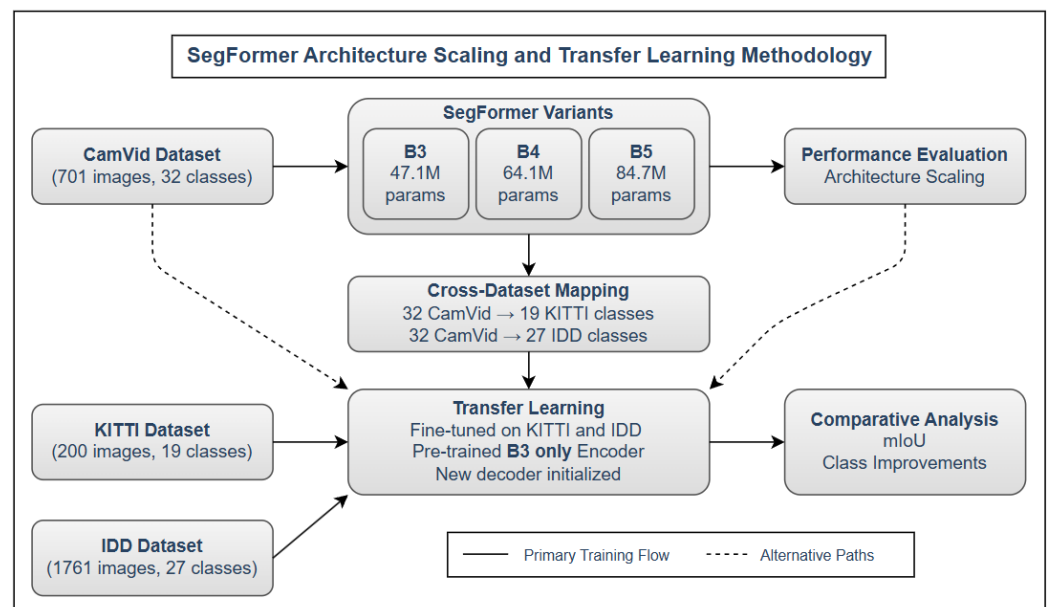


Figure 1. End-to-end pipeline: SegFormer variants are trained on CamVid, transferred to KITTI and IDD with class mappings, and interpreted via confidence heatmaps.

4.1. SegFormer Architecture Overview

SegFormer [6] consists of three core components: (i) *overlapping patch embedding* to preserve spatial details, (ii) a *hierarchical transformer encoder* that extracts multi-scale contextual

features, and (iii) a *lightweight decoder* that fuses these features via an MLP and performs semantic prediction.

We consider three SegFormer variants, B3, B4, and B5, differentiated by model size, capacity, and architectural configurations:

- **B3:** With 47.1 M parameters, this variant employs a hierarchical structure of 12 transformer layers distributed across 4 stages (2, 3, 6, 3), where each stage operates at progressively reduced spatial resolutions of $\frac{H}{4} \times \frac{W}{4}$, $\frac{H}{8} \times \frac{W}{8}$, $\frac{H}{16} \times \frac{W}{16}$, and $\frac{H}{32} \times \frac{W}{32}$. It utilizes a hidden dimension of 512 in the deeper layers and a reduction ratio of 1 in its efficient self-attention mechanism.
- **B4:** The intermediate variant contains 64.1 M parameters with 12 transformer layers arranged in a (2, 2, 8, 2) configuration across the four stages. B4 employs a larger embedding dimension of 640, increasing its capacity to model complex relationships while maintaining reasonable computational demands.
- **B5:** The largest variant with 84.7 M parameters, B5 maintains the (2, 2, 8, 2) layer distribution of B4 but expands the embedding dimension to 768. This provides substantially increased representational capacity and attention width, allowing for more nuanced feature extraction and relationship modeling.

All three variants share the same lightweight MLP decoder structure, which aggregates multi-level features from the hierarchical encoder through a simple yet effective design. For all variants, we employ the same efficient self-attention mechanism with a linear complexity of $O(N)$ instead of the standard quadratic $O(N^2)$ complexity, achieved through a sequence reduction process. This attention mechanism is mathematically expressed as

$$\text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{Q(RK)^T}{\sqrt{d}}\right)(RV) \quad (1)$$

where R represents the reduction operation that decreases the sequence length by a factor of the reduction ratio.

4.2. Cross-Dataset Transfer Learning

We use CamVid as the source domain for transfer learning due to its structured annotations and clean urban scenes. SegFormer-B3 serves as the base model for knowledge transfer. For each target dataset, KITTI and IDD, we perform the following steps:

1. Initialize the encoder using CamVid-pretrained weights.
2. Reinitialize the decoder to match the target dataset's class taxonomy.
3. Adapt input–output pipelines using custom class mappings.
4. Fine-tune the model with a reduced learning rate to avoid catastrophic forgetting.

Our mapping strategy includes: Direct Mappings (e.g., *road* → *road*), Semantic Mappings (e.g., *bicyclist* → *rider*), and Novel Classes (e.g., *autorickshaw* in IDD).

4.3. Transfer Learning Algorithm

To formalize our cross-dataset knowledge transfer approach, we present Algorithm 1, which details the process of transferring learned representations from a source urban scene dataset (CamVid) to target datasets (KITTI and IDD) with different class taxonomies and visual characteristics.

Algorithm 1 Cross-Dataset Knowledge Transfer for Semantic Segmentation.**Require:** Source dataset $\mathcal{D}_S$ with C_S classes, Target dataset $\mathcal{D}_T$ with C_T classes**Require:** Class mapping function $M : C_S \rightarrow C_T$ **Ensure:** Target-adapted model θ_T

```

1: Phase 1: Source Domain Pretraining
2: Initialize SegFormer model  $\theta_S$  with random weights
3: Train  $\theta_S$  on  $\mathcal{D}_S$  using loss  $\mathcal{L}_{\text{total}}$ 
4: Store optimized source parameters  $\theta_S^*$ 
5: Phase 2: Cross-Domain Parameter Transfer
6: Initialize target model  $\theta_T$  with encoder weights from  $\theta_S^*$ 
7: Randomly initialize decoder parameters of  $\theta_T$  for  $C_T$  classes
8: Apply class mapping  $M$  to align source and target semantics
9: Phase 3: Target Domain Fine-tuning
10: Set learning rate  $\alpha_T = 0.3 \times \alpha_S$ 
11: for epoch = 1 to max_epochs do
12:   Update  $\theta_T$  by minimizing  $\mathcal{L}_{\text{total}}$  on  $\mathcal{D}_T$ 
13:   Evaluate on validation set  $\mathcal{D}_{T,\text{val}}$ 
14:   if early stopping criterion met then
15:     break
16:   end if
17: end for
18: return optimized target model  $\theta_T$ 

```

4.4. Loss Function

To ensure robust learning across imbalanced classes and complex boundary regions, we employ a multi-component loss function that addresses various aspects of semantic segmentation quality:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda_1 \cdot \mathcal{L}_{\text{IoU}} + \lambda_2 \cdot \mathcal{L}_{\text{boundary}} \quad (2)$$

The class-weighted cross-entropy component $\mathcal{L}_{\text{CE}}$ addresses class imbalance by applying inverse frequency weighting:

$$\mathcal{L}_{\text{CE}} = - \sum_{i=1}^{H \times W} \sum_{c=1}^C w_c \cdot y_{i,c} \cdot \log(p_{i,c}) \quad (3)$$

where $y_{i,c}$ is the ground truth, $p_{i,c}$ is the predicted probability, and w_c is the class weight calculated as $w_c = \frac{f_j}{f_c}$, with f_c being the frequency of class c in the training set.

The IoU loss component $\mathcal{L}_{\text{IoU}}$ focuses on optimizing the Intersection-over-Union metric directly:

$$\mathcal{L}_{\text{IoU}} = 1 - \frac{1}{C} \sum_{c=1}^C \frac{\sum_i p_{i,c} \cdot y_{i,c}}{\sum_i (p_{i,c} + y_{i,c} - p_{i,c} \cdot y_{i,c})} \quad (4)$$

The boundary-aware component $\mathcal{L}_{\text{boundary}}$ enhances precision at class transitions:

$$\mathcal{L}_{\text{boundary}} = \frac{1}{|\mathcal{B}|} \sum_{i \in \mathcal{B}} \sum_{c=1}^C \beta_i \cdot |p_{i,c} - y_{i,c}| \quad (5)$$

where $\mathcal{B}$ is the set of boundary pixels identified using a Sobel edge detector on the ground truth, and β_i is a distance-based weight that emphasizes pixels closer to boundaries.

The coefficients $\lambda_1 = 0.4$ and $\lambda_2 = 0.8$ were determined through systematic grid search on the validation set. This configuration achieved an optimal balance between overall segmentation accuracy and boundary precision.

4.5. Explainability via Confidence Heatmaps

To interpret model behavior, we generate confidence heatmaps that visualize softmax entropy per pixel. These maps provide spatial cues on:

- High-confidence regions (well-learned objects);
- Uncertain predictions (occlusions, rare classes);
- Model confusion at boundaries.

These visualizations help reveal failure modes, guide further training improvements, and support safer model deployment in real-world scenarios.

5. Evaluation Metrics and Results

This section presents a rigorous quantitative and qualitative assessment of our SegFormer experiments across the CamVid, KITTI, and IDD datasets. We first establish the mathematical foundation of our evaluation framework, and then analyze the performance of different SegFormer variants and the effectiveness of our cross-dataset transfer learning approach.

5.1. Mathematical Formulation of Evaluation Metrics

Mean Intersection over Union (mIoU) measures the overlap between predicted and ground truth segmentation masks for each class, and then averages across all classes:

$$\text{mIoU} = \frac{1}{C} \sum_{c=1}^C \frac{TP_c}{TP_c + FP_c + FN_c} \quad (6)$$

In our CamVid experiments, $C = 32$ semantic classes, while KITTI uses $C = 19$ classes and IDD uses $C = 27$ classes.

Pixel Accuracy (PA) quantifies the overall proportion of correctly classified pixels:

$$\text{PA} = \frac{\sum_{c=1}^C TP_c}{\sum_{c=1}^C (TP_c + FN_c)} \quad (7)$$

Convergence Acceleration Factor (CAF) measures the reduction in training time:

$$\text{CAF} = \frac{\text{Epochs}_{\text{from scratch}}}{\text{Epochs}_{\text{transfer}}} \quad (8)$$

Class-Specific Transfer Gain (CSTG) quantifies the improvement for each semantic class after transfer learning:

$$\text{CSTG}_c = \frac{\text{IoU}_{c,\text{transfer}} - \text{IoU}_{c,\text{scratch}}}{\text{IoU}_{c,\text{scratch}}} \times 100\% \quad (9)$$

5.2. Architecture Scaling Analysis on CamVid

We systematically evaluated three SegFormer variants (B3, B4, and B5) on the CamVid dataset to understand the relationship between model capacity and segmentation performance. Table 2 summarizes our findings.

Table 2. Comprehensive evaluation of SegFormer variants on CamVid dataset.

Model	Params (M)	mIoU (%)	PA (%)	Inference Time (ms)
SegFormer-B3	47.1	77.9	94.3	25.3
SegFormer-B4	64.1	78.5	94.7	28.5
SegFormer-B5	84.7	82.4	95.6	32.8

The relationship between model size and performance follows a sub-linear pattern, indicating diminishing returns as model capacity increases as seen in the Figure 2. While SegFormer-B5 achieves the highest accuracy, the performance gain over B3 (+4.5% mIoU) comes at the cost of significantly increased parameters (+80%) and inference time (+29.6%).

For our cross-dataset transfer learning experiments, we selected SegFormer-B3 as the base model due to its favorable balance between accuracy and efficiency.

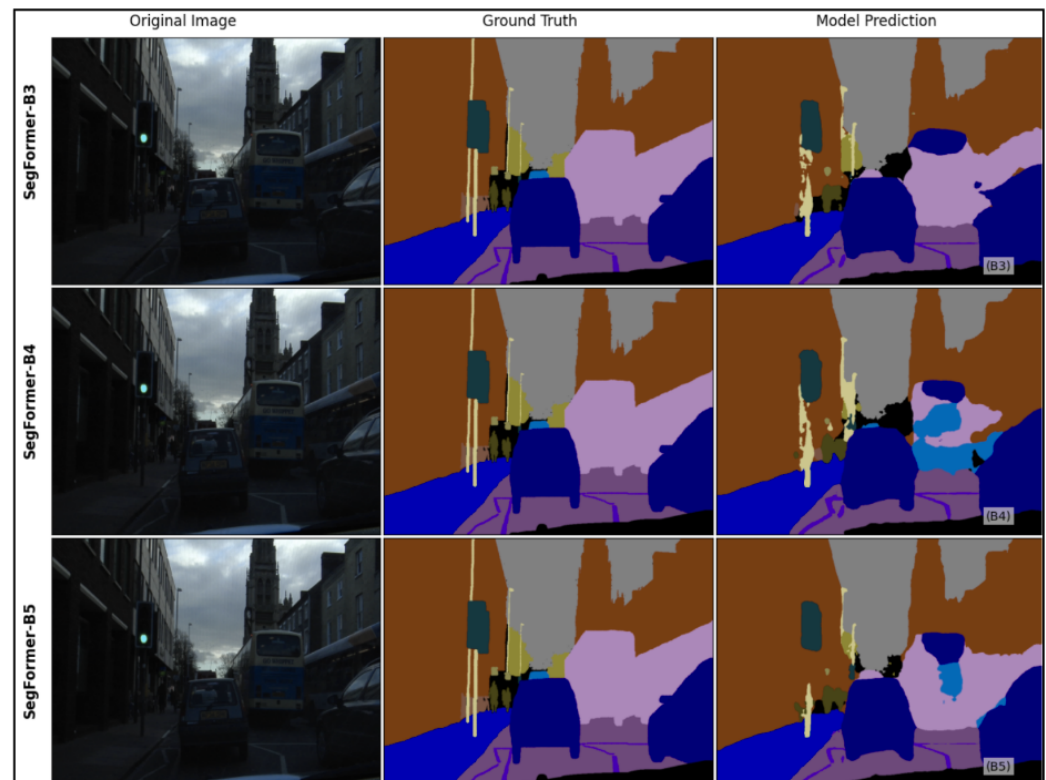


Figure 2. Qualitative segmentation results of SegFormer variants on CamVid. Larger models show improved boundary precision and semantic consistency.

5.3. KITTI Transfer Performance

We transferred knowledge from CamVid (source domain) to KITTI (target domain) by initializing a SegFormer-B3 model with weights pretrained on CamVid, and then fine-tuning on KITTI training data. Table 3 quantifies the class-specific benefits of this transfer learning approach which can also be seen in the Figure 3.

Table 3. Performance improvements on KITTI after transfer learning from CamVid.

Class	IoU (Scratch)	IoU (Transfer)	Gain (%)
Wall	0.4953	0.6476	+30.75
Sidewalk	0.4800	0.5241	+9.18
Bus	0.5421	0.5824	+7.44
Traffic Sign	0.6039	0.6353	+5.19
Bicycle	0.1100	0.1143	+3.95

The magnitude of improvement varies significantly across classes, with structural elements showing the largest gains. The overall mIoU improvement from 52.08% to 53.42% may appear modest, but this aggregate metric obscures the substantial class-specific gains. Moreover, the training efficiency improvement is dramatic, with a 61.1% reduction in

training time from 18 epochs to 7 epochs. We can see the difference between the baseline model and the improved transfer learning model in Figure 4.

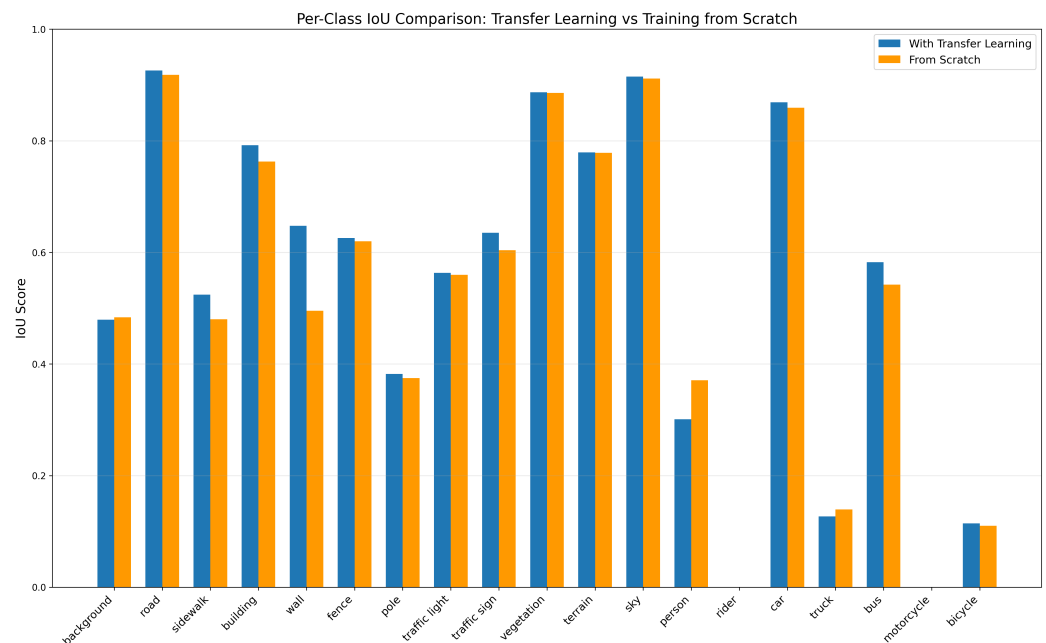


Figure 3. Per-class IoU comparison between transfer learning (blue) and training from scratch (orange) on KITTI. Classes with complex features and limited examples show the greatest relative improvements.

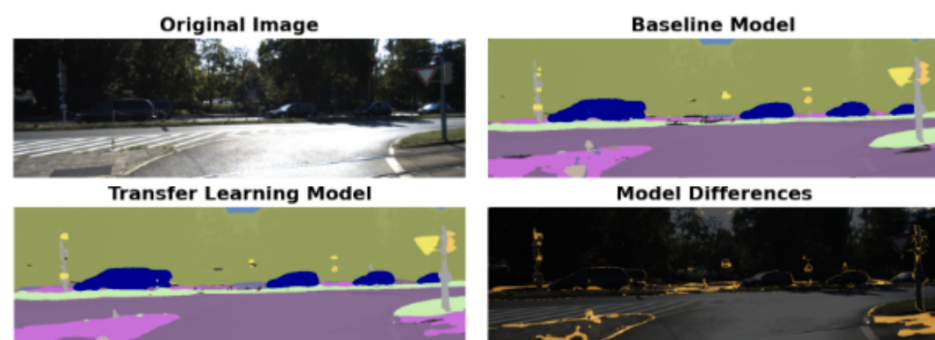


Figure 4. Qualitative results on KITTI. (Top left): original image. (Top right): baseline model prediction. (Bottom left): transfer learning model prediction. (Bottom right): visualization of model differences.

5.4. IDD Transfer Performance

The Indian Driving Dataset presents a more challenging transfer scenario due to greater visual differences from European urban scenes. Table 4 shows that transfer learning yields even larger improvements on IDD than on KITTI which can also be seen in Figure 5.

Table 4. Performance improvements on IDD after transfer learning from CamVid.

Class	IoU (Scratch)	IoU (Transfer)	Gain (%)
Motorcycle	0.1748	0.3019	+72.74
Rider	0.2156	0.3103	+43.91
Traffic Light	0.3351	0.4145	+23.69
Autotrickshaw	0.2925	0.3389	+15.87
Curb	0.3937	0.5088	+29.27

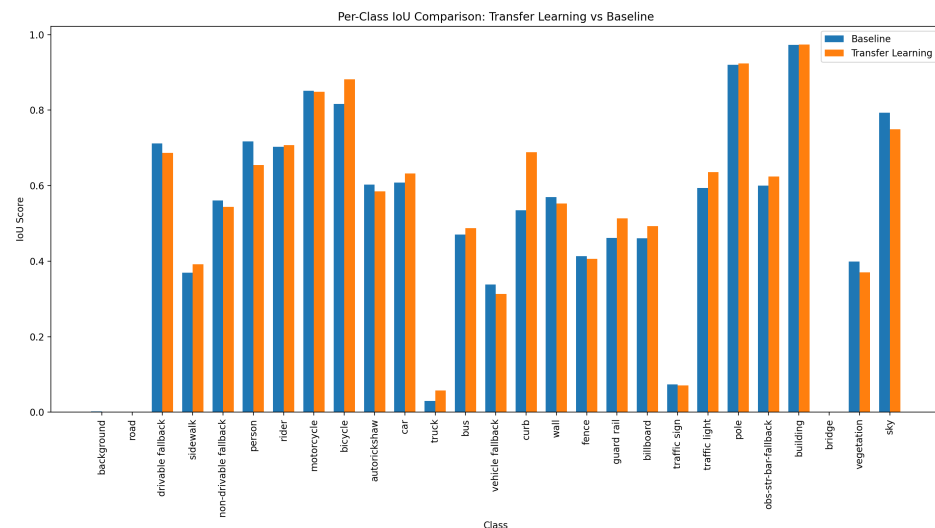


Figure 5. Per-class IoU comparison between transfer learning (blue) and training from scratch (orange) on IDD. Much larger improvements demonstrate transfer learning benefits for datasets with greater domain shift.

The remarkably high improvement for “Motorcycle” (+72.74%) can be attributed to low baseline performance due to class imbalance and visual complexity, combined with transferable features from similar classes in CamVid. Even classes unique to IDD like “Autorickshaw” benefit from transfer learning (+15.87%), demonstrating that lower-level features learned from CamVid transfer effectively despite semantic differences. We can see the difference between the baseline model and the improved transfer learning model in Figure 6.

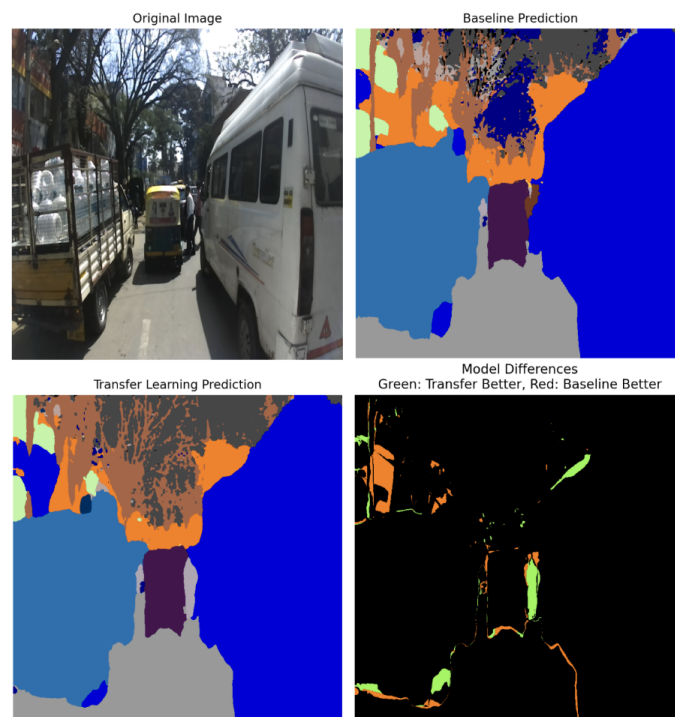


Figure 6. Qualitative results on IDD showing improved segmentation of region-specific vehicles and structures.

6. Explainability and Interpretability

While performance metrics such as mean Intersection over Union (mIoU) and pixel-wise accuracy provide essential quantitative benchmarks for evaluating semantic segmentation models, they often fall short in explaining the underlying reasoning behind model predictions. To bridge this interpretability gap, we incorporate two complementary techniques: confidence heatmaps and Gradient-weighted Class Activation Mapping (Grad-CAM).

Confidence heatmaps serve as a visual indicator of the model's certainty in its predictions. For each pixel, the confidence score is defined as the maximum softmax probability across all classes:

$$\text{Confidence}(i) = \max_{c \in \{1, \dots, C\}} p_{ic} \quad (10)$$

This scalar value reflects the model's belief in its most likely prediction at that location. As illustrated in Figure 7, high-confidence regions typically correspond to well-represented classes like roads and vehicles, while low-confidence regions often cluster around ambiguous areas, class boundaries, and occluded objects.

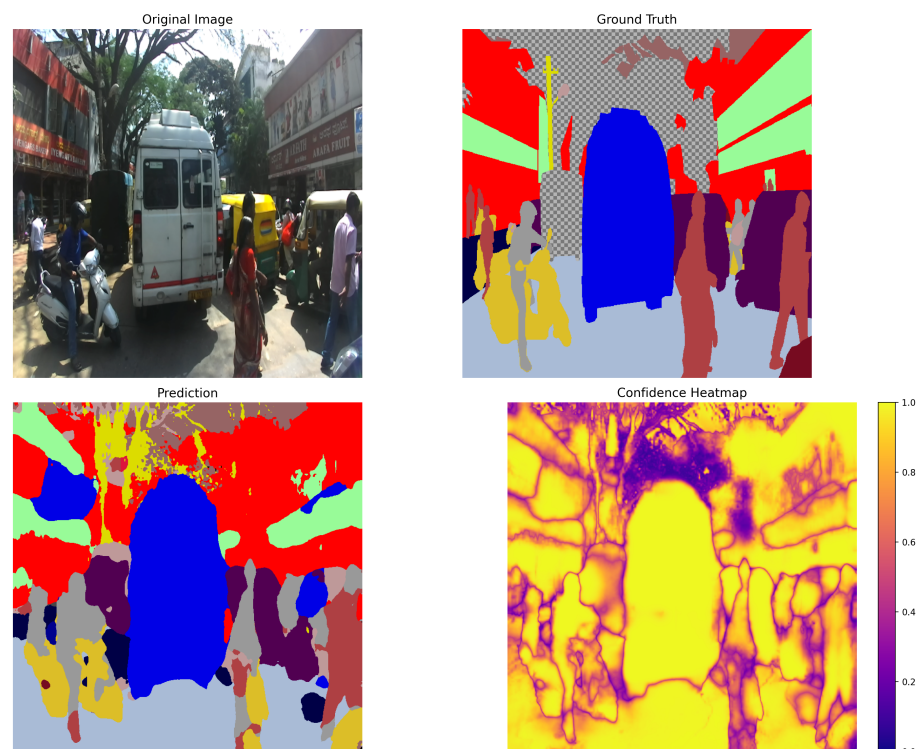


Figure 7. Confidence heatmap on IDD dataset: Brightness indicates model certainty. Lower confidence occurs near class boundaries and occlusions.

Grad-CAM is employed to generate class-specific activation maps, enabling us to probe the model's internal reasoning. This technique computes the gradient of the class-specific logit y^c with respect to the activation maps A^k of the final convolutional layer. The final Grad-CAM heatmap $L_{\text{Grad-CAM}}^c$ for class c is obtained by

$$L_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c A^k \right) \quad (11)$$

where α_k^c are importance weights for each channel. As shown in Figure 8, each class triggers distinct regions of activation, with road class activations across pavement areas and pole class activations in vertical regions.

In our implementation, Grad-CAM is applied to the final encoder feature representation prior to decoder fusion, enabling visualization of spatial attention within the transformer backbone while preserving the segmentation-specific prediction pipeline.

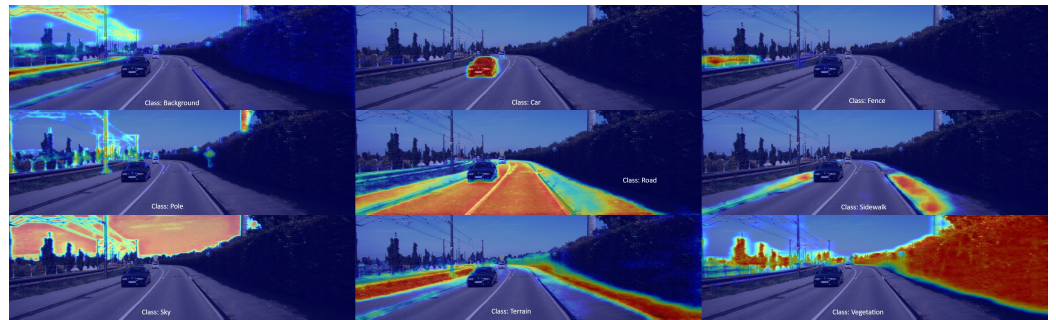


Figure 8. Grad-CAM visualizations on KITTI dataset: strong class-specific activations in structured road environments, with clear focus on roads, cars, sidewalks, and poles.

To examine generalizability across diverse driving environments, we evaluate Grad-CAM outputs on the CamVid and IDD datasets. The CamVid dataset features dense urban traffic with frequent pedestrian and bicycle interactions. As illustrated in Figure 9, the model shows broader, more diffused attention around dynamic objects, while stationary classes maintain sharp activations.

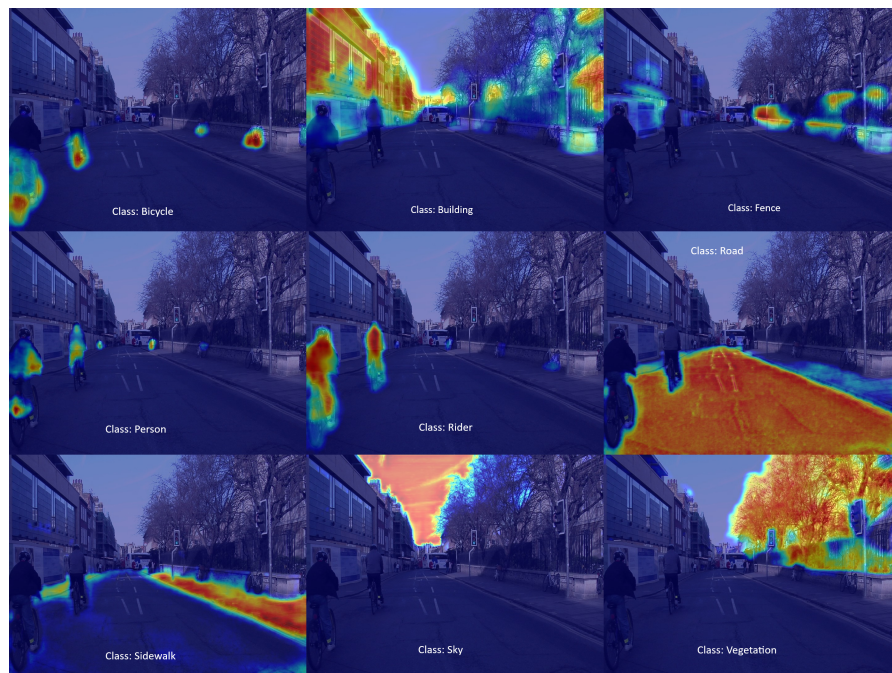


Figure 9. Grad-CAM visualizations on CamVid dataset: Attention maps for dynamic objects are more diffused, while static classes are sharply localized.

The IDD dataset poses greater challenges due to unstructured Indian street scenes. As shown in Figure 10, attention remains strong for high-frequency classes like road and vegetation, but becomes scattered for occluded classes, reflecting increased model uncertainty in complex contexts.

Visual Results with Prediction Overlay

To supplement our quantitative metrics and interpretability analyses, we present qualitative visualizations demonstrating SegFormer-B3's segmentation capabilities across diverse real-world driving environments.

Figure 11 shows results from the CamVid dataset, where the model demonstrates strong performance in delineating core urban classes such as roads, sidewalks, and buildings, as well as dynamic objects like cyclists and pedestrians.

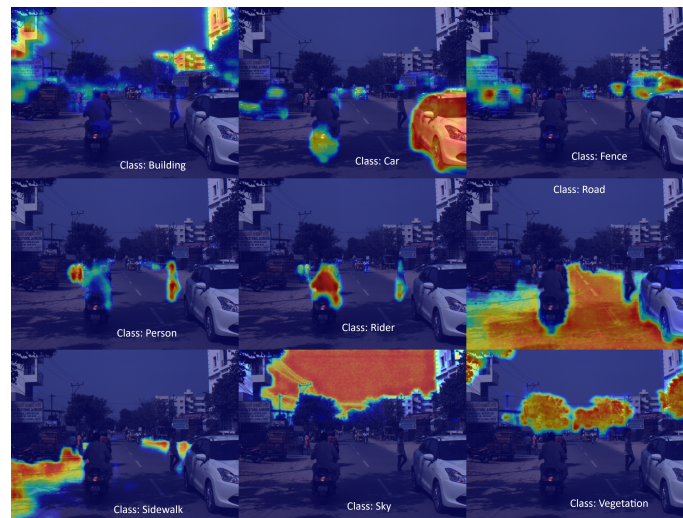


Figure 10. Grad-CAM visualizations on IDD dataset: strong attention for road and vegetation; scattered focus for occluded classes like rider and person.



Figure 11. Prediction overlay on CamVid dataset: from (left) to (right)—original input, ground truth, and predicted segmentation. Strong spatial alignment for roads, buildings, and dynamic classes.

Figure 12 shows the results of the KITTI dataset, featuring structured suburban driving scenes. SegFormer-B3 exhibits precise segmentation of road surfaces, curbs, and sidewalks, with clear improvements from transfer learning.

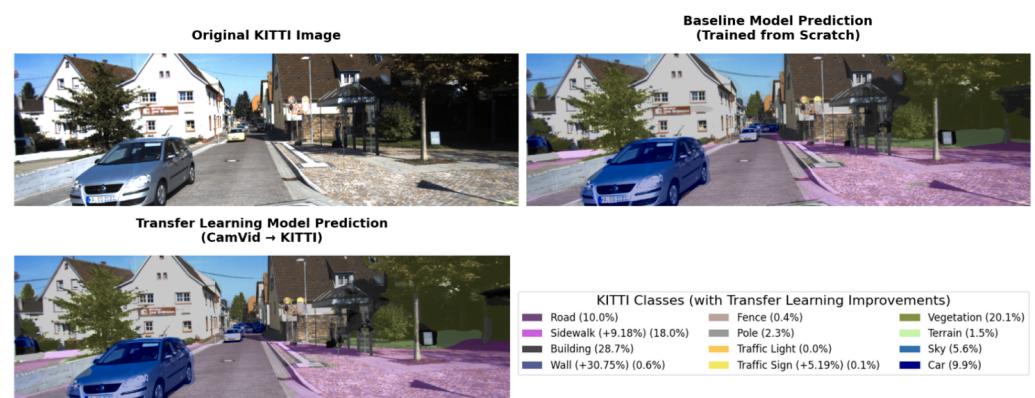


Figure 12. Prediction overlays on KITTI dataset. (Top image): baseline model; (bottom image): transfer learning from CamVid. Clear improvements in road structure delineation.

Figure 13 presents segmentation examples from the IDD dataset with unstructured Indian street scenes. Despite complexity, SegFormer-B3 accurately segments roads and vehicles, and adapts well to underrepresented classes after transfer learning.



Figure 13. Prediction overlays on IDD dataset. (Top image): baseline prediction; (bottom image): transfer learning from CamVid. Improved detection of road area, vegetation, and person classes.

These visualizations collectively illustrate that SegFormer-B3 generalizes well across varying spatial layouts and lighting conditions, demonstrating its suitability for real-time, resource-constrained environments like autonomous vehicles.

7. Conclusions and Future Work

This study presented a comprehensive evaluation of SegFormer for urban scene segmentation, contributing significant insights in three key dimensions: architecture scaling, cross-dataset transfer learning, and model interpretability. Rather than introducing a new architecture, our contribution lies in providing an integrated empirical analysis that combines scaling behavior, cross-dataset transfer, and explainability within a unified evaluation framework.

Our architectural scaling experiments on the CamVid dataset demonstrated that while SegFormer-B5 achieves the highest accuracy at 82.4% mIoU, the efficiency–performance trade-off favors the more balanced SegFormer-B3 variant for practical deployments. The relationship between model capacity and performance follows a sub-linear pattern, with diminishing returns as parameter count increases, an important consideration for resource-constrained autonomous systems.

Our cross-dataset transfer learning investigation revealed that knowledge acquired from CamVid can significantly enhance segmentation performance in both structurally similar environments (KITTI) and substantially different driving contexts (IDD). Transfer benefits were particularly pronounced for classes with limited training examples and complex geometric structures, with improvements as high as 30.75% for “Wall” in KITTI and 72.74% for “Motorcycle” in IDD. Moreover, transfer learning dramatically reduced training time by 61.1% on KITTI, demonstrating considerable practical value for model adaptation across geographic regions.

Our interpretability analysis incorporated both confidence heatmaps and Gradient-weighted Class Activation Mapping (Grad-CAM) to enhance transparency in the model's internal decision-making. Confidence heatmaps allowed us to visualize model certainty at the pixel level, while Grad-CAM provided class-specific saliency visualizations. This layer of explainability is critical for safety-critical applications such as autonomous driving, where understanding the rationale behind predictions is essential for trust and deployment readiness. These multifaceted contributions offer significant value for real-world autonomous systems facing the challenges of varied deployment conditions, limited labeled data availability, and model explainability requirements. By demonstrating SegFormer's adaptability across diverse urban environments, our work provides a foundation for developing more geographically robust perception systems. Although this study focuses on image-domain transformer models, extending similar transferability and explainability principles to 3D point cloud representations remains an important future direction. Recent work exploring physics-aware point cloud learning suggests that integrating geometric reasoning with transformer-based perception could further enhance robustness in complex urban environments. Investigating how image-domain transfer strategies translate to multimodal or 3D spatial representations is therefore a promising avenue for future research.

Several promising directions extend from our current findings. First, exploring multi-source pretraining and domain adaptation strategies could further enhance model generalization. Rather than transferring from a single source dataset, simultaneously leveraging knowledge from multiple geographically diverse datasets could create more universal feature representations.

Second, evaluating SegFormer deployment on embedded hardware platforms would address real-world implementation challenges. Quantifying the latency–accuracy trade-offs across different computational constraints would provide practical guidelines for autonomous vehicle manufacturers and smart city developers.

By pursuing these research directions, we anticipate significant advancements in the development and deployment of efficient, transferable, and interpretable segmentation models for understanding in diverse global environments of urban scenes.

Author Contributions: S.B.A. conceptualized the study and supervised the research. T.S.H. and A.P. developed the proposed methodology and implemented the experiments. T.S.H. curated the dataset and performed preprocessing. T.S.H. and A.P. conducted data analysis and interpretation of the results. S.B.A. reviewed the results and provided critical revisions. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The datasets analyzed during the current study are publicly available. The implementation of our proposed model can be accessed on Github: <https://github.com/mrw0nd3rfu1/Evaluating-Architecture-Scalability-in-Urban-Scene-Segmentation-Using-Explainable-AI/blob/main/README.md> (accessed on 1 January 2026).

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

References

1. Badrinarayanan, V.; Kendall, A.; Cipolla, R. SegNet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 2481–2495. [CrossRef]
2. Jégou, S.; Drozdal, M.; Vazquez, D.; Romero, A.; Bengio, Y. The One Hundred Layers Tiramisu: Fully convolutional DenseNets for semantic segmentation. In Proceedings of the IEEE CVPR Workshops, Honolulu, HI, USA, 21–26 July 2017; pp. 11–19.

3. Yu, C.; Wang, J.; Peng, C.; Gao, C.; Yu, G.; Sang, N. BiSeNet: Bilateral segmentation network for real-time semantic segmentation. In Proceedings of the ECCV, Munich, Germany, 8–14 September 2018.
4. Xu, X.; Li, Y.; Wu, B.; Yang, W. PIDNet: A real-time semantic segmentation network inspired by PID controllers. In Proceedings of the ECCV, Tel Aviv, Israel, 23–27 October 2022.
5. Zhang, Y.; Li, K.; Chen, K.; Wang, X.; Hu, J. RTFormer: Efficient design for real-time semantic segmentation with transformer. In Proceedings of the CVPR, New Orleans, LA, USA, 18–24 June 2022.
6. Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J.M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with transformers. In Proceedings of the Advances in Neural Information Processing Systems, Virtual, 6–14 December 2021; Volume 34, pp. 12077–12090.
7. Lombardi, A.; Nardo, E.D.; Ciaramella, A. Skip-SegFormer Efficient Semantic Segmentation for urban driving. In Proceedings of the Ital-IA 2023: 3rd National Conference on Artificial Intelligence, Pisa, Italy, 29–31 May 2023; pp. 54–59.
8. Zhao, L.; Wei, X.; Chen, J. CFF-SegFormer: Lightweight network modeling based on SegFormer. *IEEE Access* **2023**, *11*, 84372–84384.
9. Hoyer, L.; Dai, D.; Gool, L.V. DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation. In Proceedings of the CVPR, New Orleans, LA, USA, 18–24 June 2022; pp. 10157–10167.
10. Kendall, A.; Badrinarayanan, V.; Cipolla, R. Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. In Proceedings of the British Machine Vision Conference (BMVC), York, UK, 19–22 September 2016.
11. Brostow, G.J.; Fauqueur, J.; Cipolla, R. Semantic object classes in video: A high-definition ground truth database. *Pattern Recognit. Lett.* **2009**, *30*, 88–97. [[CrossRef](#)]
12. Geiger, A.; Lenz, P.; Stiller, C.; Urtasun, R. Vision meets robotics: The KITTI dataset. *Int. J. Robot. Res.* **2013**, *32*, 1231–1237. [[CrossRef](#)]

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