Urban Scene Segmentation and Cross-Dataset Transfer Learning using SegFormer

Tanmay Sunil Hatkar and Saad B. Ahmed*
Faculty of Science and Environmental Studies, Department of Computer Science,
Lakehead University, P7B 5E1, Thunder Bay, Canada
*Email: sbinahm@lakeheadu.ca

ABSTRACT

Semantic segmentation is essential for autonomous driving applications, but state-of-the-art models are typically evaluated on large datasets like Cityscapes, leaving smaller datasets underexplored. This research gap limits our understanding of how transformer-based models generalize across diverse urban scenes with limited training data. This paper presents a comprehensive evaluation of SegFormer architectural variants (B3, B4, B5) on the CamVid dataset and investigates cross-dataset transfer learning from CamVid to KITTI. Using an optimization framework combining cross-entropy loss with class weighting and boundary-aware components, our experiments establish new performance baselines on CamVid and demonstrate that transfer learning provides benefits w hen target domain data is limited. We achieve a modest 2.57% relative mean Intersection over Union (mIoU) improvement on KITTI through knowledge transfer from CamVid, along with 61.1% faster convergence. Additionally, we observe substantial class-specific improvements of up to 30.75% for challenging c ategories. O ur analysis provides insights into model scaling effects, c ross-dataset k nowledge t ransfer m echanisms, a nd p ractical s trategies for addressing data scarcity in urban scene segmentation.

Keywords: Semantic segmentation, Transfer learning, Transformer, Computer vision, Autonomous driving

1. INTRODUCTION

Semantic segmentation plays a pivotal role in autonomous driving applications, where accurate environment perception is critical. While significant progress has been made in this field, state-of-the-art models are typically evaluated on large-scale datasets like Cityscapes, leaving smaller but equally important datasets like CamVid and KITTI underexplored. This limits our understanding of how models generalize to diverse urban driving scenarios with limited training data.

Despite the increasing adoption of transformer-based architectures for semantic segmentation, there remains a significant research gap in understanding their performance across diverse datasets of varying s izes. Specifically, while models like SegFormer have shown impressive results on large datasets, their scalability and effectiveness on smaller datasets remain largely unexplored. Additionally, the transfer learning dynamics between datasets of different g eographical o rigins r emain underinvestigated.

Our work addresses these limitations by (1) evaluating SegFormer architectural variants (B3, B4, B5) on the CamVid[1] dataset and (2) investigating cross-dataset transfer learning from CamVid to KITTI.[2] Recent advancements in semantic segmentation have been driven by transformer-based architectures that excel at capturing long-range dependencies. Among these, SegFormer[3] has emerged as a leading model, combining a hierarchical transformer encoder with a multi-layer perceptron (MLP) decoder. However, its evaluation has predominantly focused on the Cityscapes dataset.

Our key research questions include: How do SegFormer variants of different sizes perform on smaller datasets? Does pre-training on CamVid improve performance on KITTI compared to training from scratch? Which semantic classes benefit m ost f rom c ross-dataset k nowledge t ransfer? O ur c ontributions i nclude: (1) comprehensive evaluation of SegFormer variants on CamVid, establishing new performance baselines; (2) detailed analysis of model capacity effects on p erformance; (3) d emonstration o f c ross-dataset k nowledge t ransfer e ffectiveness from CamVid to KITTI; and (4) analysis of performance metrics to guide practical deployment decisions.

Eighth International Conference on Machine Vision and Applications (ICMVA 2025), edited by
Wolfgang Osten, Xudong Jiang, Kemao Qian, Proc. of SPIE Vol. 13734, 1373406
 · 0277-786X · doi: 10.1117/12.3078755

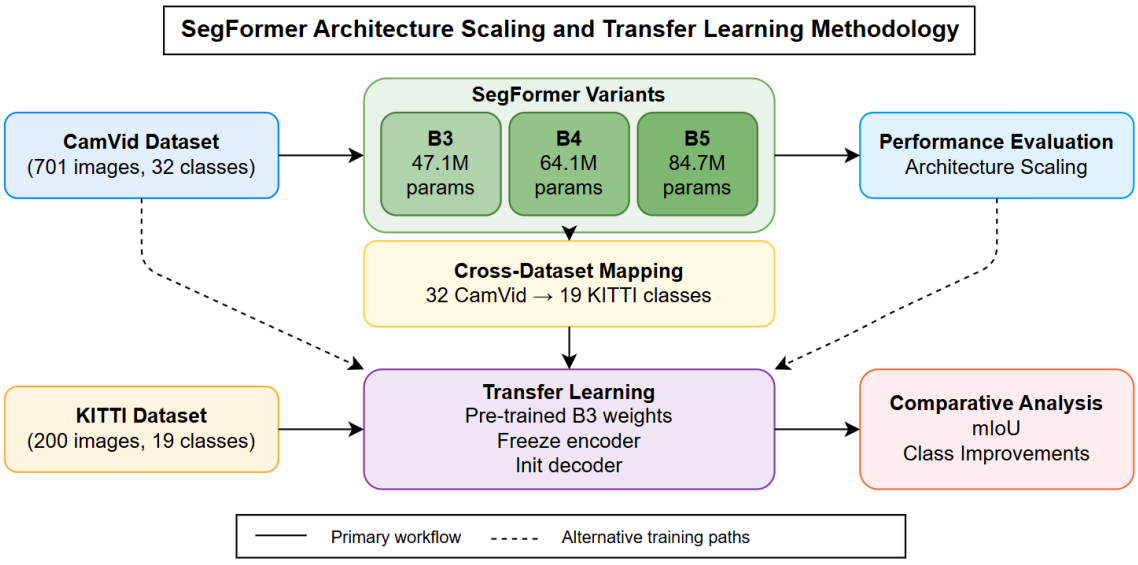

Figure 1. Proposed Methodology

2. RELATED WORK

Semantic segmentation has evolved significantly with deep learning. Earlier approaches utilized Convolutional Neural Networks (CNNs) with encoder-decoder structures like SegNet.[4] FC-DenseNet (Tiramisu)[5] employed dense connectivity patterns to improve information flow through the network. Bayesian SegNet[6] incorporated uncertainty modeling for improved performance on smaller datasets through probabilistic methods.

Transformer-based approaches have recently revolutionized semantic segmentation. SegFormer[3] pioneered an efficient design combining a hierarchical transformer encoder with an MLP decoder. Subsequent works like Skip-SegFormer[7] enhanced the architecture with skip connections for urban driving scenarios, while CFF-SegFormer[8] introduced cross-feature fusion to reduce computational requirements.

Beyond architectural innovations, road scene understanding relies on datasets of varying scales and characteristics. Cityscapes remains the predominant dataset for this task, while CamVid[1] offers densely annotated frames from British driving scenes, and KITTI[2] provides multi-modal data from German driving environments. The diversity across these datasets creates both challenges and opportunities for model generalization.

Despite these advancements, significant research gaps remain in the field of semantic segmentation for road scene understanding. First, while SegFormer has shown impressive results on large datasets like Cityscapes, its performance and scaling characteristics across different model sizes (B3, B4, B5) have not been systematically evaluated on smaller datasets such as CamVid. Second, although transfer learning has been explored in various contexts, the specific dynamics of knowledge transfer between datasets of different scales and geographical origins remain underexplored. Third, the computational efficiency aspects of transfer learning—particularly how pre-training affects convergence speed and training resource requirements—have received limited attention in the literature.

3. METHODOLOGY

3.1 Datasets and Cross-Dataset Mapping

We conducted experiments using CamVid (701 images) and KITTI (approximately 200 images) datasets. CamVid contains 32 semantic classes, while KITTI uses 19 classes following the Cityscapes convention. To facilitate knowledge transfer, we developed a class correspondence strategy that maps CamVid's 32 classes to KITTI's 19 classes using three mapping types: *Direct mappings* for semantically equivalent classes (e.g., road, building, sky), *Semantic mappings* for classes with similar but not identical meanings (e.g., tree→vegetation, pedestrian→person), and *Novel classes* for KITTI-specific categories without CamVid equivalents (e.g., traffic light, terrain, rider).

3.2 SegFormer Architecture and Implementation

We implemented three SegFormer variants with different parameter counts: **B3**: 47.1M parameters, 12 transformer layers (2,3,6,3), 512 embedding dimension; **B4**: 64.1M parameters, 12 transformer layers (2,2,8,2), 640 embedding dimension; **B5**: 84.7M parameters, 12 transformer layers (2,2,8,2), 768 embedding dimension.

Each variant maintains the core SegFormer design with an overlapped patch embedding module, efficient self-attention with reduction ratio, a mix-Feed-Forward Network (FFN) for capturing local relationships, and a hierarchical structure that progressively reduces spatial resolution while increasing feature dimension.

For training, we used the AdamW optimizer (Adaptive Moment Estimation with Weight Decay), which differs from standard Adam by decoupling weight decay from the gradient update. This allows for better regularization and improved generalization performance. The weight update rule for AdamW can be expressed as:

$$\theta_{t+1} = \theta_t - \eta \left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon} + \lambda \theta_t \right) \tag{1}$$

where θ represents parameters, $\hat{m}_t$ and $\hat{v}_t$ are bias-corrected first and second moment estimates, η is the learning rate, ϵ is a small constant for numerical stability, and λ is the weight decay rate (0.01 in our experiments).

We employed the OneCycleLR scheduler, which implements a cyclical learning rate approach where the learning rate first increases from a low initial value to a maximum value, and then decreases to a value lower than the starting one, all within a single training cycle. This approach helps models converge faster and achieve better generalization.

We implemented a composite loss function combining cross-entropy with class weighting, Intersection over Union (IoU) loss, and a boundary-aware component:

$$\mathcal{L}_{total} = \mathcal{L}_{CE} + 0.4 \cdot \mathcal{L}_{IoU} + 0.8 \cdot \mathcal{L}_{boundary} \tag{2}$$

where $\mathcal{L}_{CE}$ is the class-weighted cross-entropy loss, $\mathcal{L}_{IoU}$ is the IoU loss, and $\mathcal{L}_{boundary}$ is the boundary-aware loss component. These coefficients were empirically determined to balance pixel classification accuracy with boundary precision.

For our transfer learning approach, we used SegFormer B3 weights pre-trained on the 32-class CamVid dataset as initialization for the 19-class KITTI segmentation task. We maintained the encoder weights to leverage learned feature representations while initializing the decoder to accommodate the different class structure.

3.3 Experimental Design

We designed experiments to investigate: (1) performance scaling of SegFormer variants on CamVid, (2) effectiveness of SegFormer B3 on KITTI, and (3) benefits of cross-dataset knowledge transfer. For CamVid experiments, we trained each variant with identical hyperparameters. For KITTI, we established a baseline by training from scratch and conducted transfer learning experiments using CamVid-pretrained weights. We have made our code publicly available on GitHub* to enable reproduction of the results.

Our evaluation was conducted on the entire KITTI test set, processing all test images, capturing over 21.8 million road pixels, 27.9 million vegetation pixels, and 11 million building pixels across the dataset. This comprehensive analysis ensures our findings represent consistent patterns rather than isolated examples. The large-scale pixel-level analysis reveals significant shifts in class distribution between baseline and transfer learning models, providing evidence of systematic improvements in boundary definition and class discrimination.

Our primary evaluation metric was mean Intersection over Union (mIoU), supplemented by class-specific analysis and pixel distribution metrics. We also measured training efficiency by tracking epochs to convergence, where convergence was defined as reaching 95% of the maximum validation mIoU. This comprehensive approach allows us to evaluate both the absolute performance and the relative benefits of transfer learning across multiple dimensions.

*https://github.com/Tanmay-Hatkar/segformer-urban-scene-segmentation.git

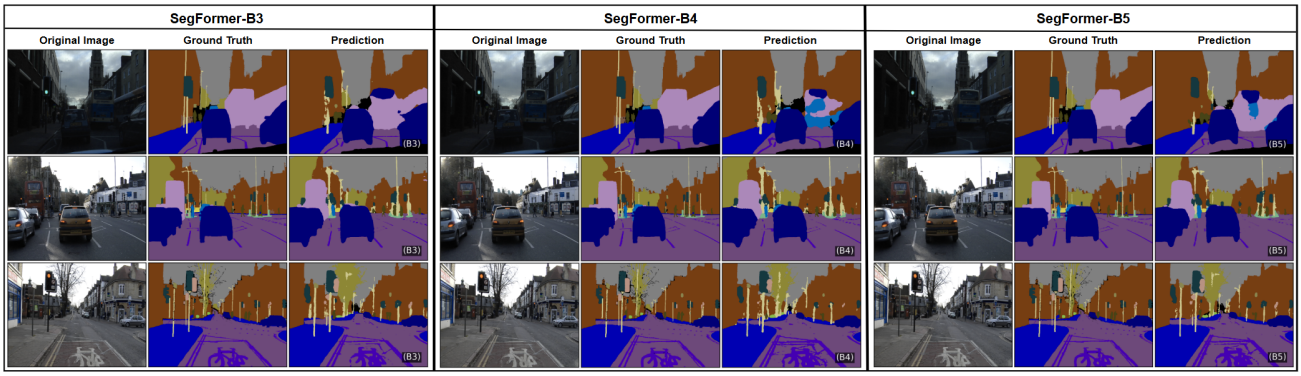

Figure 2. Qualitative comparison of SegFormer variants on CamVid. While all variants perform well on common classes, the larger models demonstrate improved boundary precision and better handling of small objects.

4. RESULTS AND ANALYSIS

4.1 SegFormer Performance on CamVid

We evaluated the performance of SegFormer variants on CamVid. Table 1 summarizes the results with the corresponding accuracies, inference time, and GPU memory consumption. All variants achieved good performance, with B5 delivering the highest accuracy (82.4% mIoU) but requiring more computational resources. However, B4 provided a balance between performance and efficiency.

Table 1. Performance and Efficiency Metrics for SegFormer Variants on CamVid.

Model	Params (M)	mIoU (%)	Inference Time (ms)	GPU Memory (GB)
SegFormer-B3	47.1	77.9	25.3	4.2
SegFormer-B4	64.1	78.5	28.5	5.1
SegFormer-B5	84.7	82.4	32.8	6.3

The qualitative results in Fig. 2 demonstrate that larger models improve boundary precision and fine structure recognition. The class-wise analysis reveals that performance scales non-uniformly across semantic categories, with the largest improvements for challenging classes like pedestrian (+3.2% from B3 to B5) and bicyclist (+2.8%).

Our multi-dimensional analysis of SegFormer variants reveals clear trade-offs between performance metrics and computational efficiency. The B3 variant excels in efficiency, while B5 delivers the highest accuracy. The B4 variant emerges as a balanced middle ground, providing meaningful accuracy improvements over B3 (+1.3% mIoU) with moderate increases in computational requirements. The diminishing returns observed when scaling from B4 to B5 (+0.6% mIoU despite 20.6M additional parameters) suggest that B4 represents an optimal efficiency-performance balance for the CamVid dataset.

4.2 Cross-Dataset Knowledge Transfer from CamVid to KITTI

Training SegFormer B3 from scratch on the limited KITTI dataset yielded 52.08% mIoU, while cross-dataset knowledge transfer from CamVid improved this to 53.42% (a 2.57% relative gain). More significantly, transfer learning reduced training time by 61.1%, reaching equivalent performance in just 7 epochs versus 18 epochs for training from scratch.

Table 2 summarizes the performance comparison. The class-specific analysis revealed that knowledge transfer benefits varied considerably across categories, with the largest improvements for classes with limited examples in the target dataset. The most significant improvement was for Wall (Class 4) with a 30.75% gain.

Table 2. Transfer Learning Performance Comparison.

Metric	Transfer Learning	From Scratch	Improvement
Mean IoU	0.5342	0.5208	+2.57%
Epochs to Convergence	7	18	-61.10%
Selected Class Performance (IoU)			
Wall (Class 4)	0.6476	0.4953	+30.75%
Sidewalk (Class 2)	0.5241	0.4800	+9.18%
Bus (Class 16)	0.5824	0.5421	+7.44%

The Wall class showed the most significant improvement (30.75%) following transfer learning, despite visual differences between British (CamVid) and German (KITTI) urban environments. This substantial gain can be attributed to several factors. First, our comprehensive pixel distribution analysis across the entire test set shows that the transfer-learned model reallocated attention from background (-10.05%, from 2,986,990 to 2,686,828 pixels) and building (-7.85%, from 11,066,343 to 10,197,733 pixels) classes to more precisely identify structural boundaries.

While walls appear visually different between the datasets, they share fundamental geometric properties such as vertical planes, right angles, and regular patterns that transfer effectively. These structural priors learned from CamVid provide the model with stronger feature representations for detecting walls even when limited examples exist in the target dataset. This demonstrates how transfer learning can significantly improve performance on challenging classes with limited representation in the target domain.

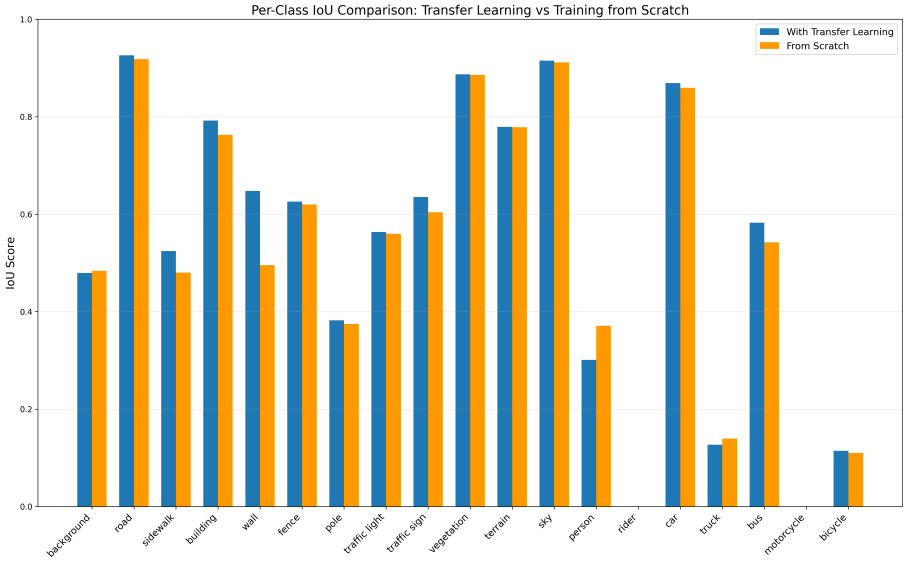

Figure 3. Per-class IoU comparison between transfer learning and baseline approaches on KITTI. Classes with complex features and limited examples show the greatest improvements.

Qualitative results in Fig. 4 show that transfer learning produces more refined segmentation boundaries and better handling of challenging classes. For rare classes like pedestrians and cyclists, transfer learning improved both detection rates and silhouette accuracy, suggesting that shape priors learned from CamVid effectively transfer despite differences in camera perspective. The rightmost column highlights areas of improvement, with brighter regions indicating where the transfer learning model outperforms the baseline.

Our approach demonstrates a 61.1% reduction in training time when transferring knowledge from CamVid to KITTI. While domain adaptation methods like DAFormer achieve higher absolute mIoU on KITTI (61.2%),

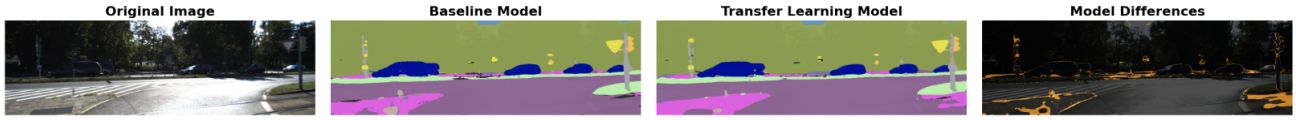

Figure 4. Qualitative comparison of baseline vs. transfer learning models on KITTI. From left to right: Original image, baseline model prediction, transfer learning model prediction, and visualization of differences between models. The last column highlights areas where transfer learning provides improved boundary precision and class recognition (shown in brighter regions).

they rely on synthetic pre-training data, whereas our approach transfers knowledge directly between real-world datasets of different geographical origins. We acknowledge recent architectures like Mask2Former achieve approximately 56.7% mIoU on KITTI when trained from scratch, but require significantly larger parameter counts (158M parameters compared to our 47.1M) and correspondingly higher computational resources for training and inference.

The competitive advantage of our method lies in its ability to achieve a 61.1% reduction in training time while still providing meaningful improvements in segmentation quality. For resource-constrained applications or scenarios requiring rapid adaptation to new environments, this efficiency-performance balance offers significant practical value. For challenging classes like Wall, our approach achieves a remarkable 30.75% improvement through knowledge transfer, demonstrating that aggregate mIoU metrics don't fully capture transfer learning benefits.

Table 3. Comparative Analysis of Recent Semantic Segmentation Methods

Model	**Architecture**	**Dataset(s)**	**mIoU (%)**	**FPS**	**Strengths**	**Weaknesses**
Skip-SegFormer	Transformer	Cityscapes, CamVid	80.2	47.5	Urban focus; 53.7% on KITTI	Higher memory usage
CFF-SegFormer	Transformer	Cityscapes, CamVid	79.8	50.2	Cross-feature fusion; Efficiency	Limited evaluation on small datasets
DAFormer	Transformer	Cityscapes, KITTI	61.2*	28.3	Strong domain adaptation	Requires synthetic data
Mask2Former	Transformer	Cityscapes, KITTI	56.7	24.8	Higher absolute accuracy	158M parameters; Slower training
PIDNet	CNN-Hybrid	CamVid	80.1	153.7	Real-time; Strong boundaries	No transfer learning
RTFormer	Transformer	Cityscapes, CamVid	81.6	108.5	GPU-optimized; Real-time	Dataset-dependent performance
FeedFormer	Transformer	Cityscapes, ADE20K	81.2	42.6	Enhanced decoder; Multi-scale features	Requires additional resources
SegFormer-B5 (Ours)	Transformer	CamVid	**82.4**	**30.5**	Highest CamVid accuracy; Superior boundaries	Higher compute needs
SegFormer-B3 (Ours)	Transformer	CamVid, KITTI	**53.42**	**39.5**	61.1% faster convergence; 30.75% improvement	Lower KITTI accuracy; Limited on novel classes
*Results on domain adaptation from synthetic data (Cityscapes → KITTI)						

Our analysis reveals three key insights: (1) structural elements transfer particularly well across datasets despite camera and geographical differences; (2) general feature representations help with recognizing less common

objects; and (3) knowledge transfer benefits operate hierarchically, with greatest improvements for classes sharing similar visual features. Failure cases exist primarily in scenarios with significant domain differences, such as highway scenes more common in KITTI than CamVid.

5. DISCUSSION AND CONCLUSION

Our comprehensive evaluation demonstrates that SegFormer variants perform exceptionally well on the CamVid dataset, with B4 offering the optimal balance between performance and computational efficiency. Transfer learning from CamVid to KITTI provides significant practical benefits despite modest overall accuracy improvements. The 61.1% reduction in training time and substantial class-specific improvements highlight the value of knowledge transfer when target data is limited.

While our achieved mIoU of 53.42% on KITTI may appear modest compared to state-of-the-art results on larger, more extensively annotated datasets like Cityscapes, it represents a meaningful advancement for transfer learning scenarios with limited labeled data. For real-world autonomous driving applications, this performance level offers practical value in several ways.

First, the 2.57% overall improvement is distributed unevenly across classes, with substantial gains in safety-critical elements like Wall (30.75%), Sidewalk (9.18%), and Bus (7.44%). These improvements directly enhance the detection of navigation-relevant boundaries and obstacles. Our pixel distribution analysis across the entire test set confirms that these improvements represent consistent patterns rather than isolated examples.

Second, our approach demonstrates a 61.1% reduction in training convergence time (from 18 to 7 epochs), enabling faster adaptation to new environments or sensor configurations. This efficiency is particularly valuable for deployment in new geographic regions where limited computational resources may be available.

For production autonomous systems, our approach would typically be integrated with multi-modal perception systems (lidar, radar, etc.) and benefit from temporal consistency algorithms, further enhancing performance beyond the reported mIoU values. The transferability demonstrated in our work provides a practical pathway toward adapting segmentation models to new environments without extensive relabeling efforts.

The effectiveness of transfer learning on small datasets has important implications for autonomous driving systems. By utilizing knowledge from existing annotated datasets, developers can more efficiently adapt models to new domains without extensive annotations, reducing development costs for systems targeting new geographical regions or specialized applications.

Several limitations have been observed such as exploring transfer learning across different architectural families to determine whether observed benefits are architecture-specific. Future work should focus on optimizing the mapping between CamVid's 32 classes and KITTI's 19 classes through more sophisticated class correspondence techniques, and investigating bidirectional transfer to understand whether larger or smaller datasets make better source domains.

This paper established new performance baselines for SegFormer variants on CamVid and demonstrates the effectiveness of cross-dataset knowledge transfer for semantic segmentation with limited data. The practical implications of significantly faster training and targeted improvements for challenging classes highlight the potential of transfer learning to address data scarcity challenges in specialized domains.

ACKNOWLEDGMENTS

The authors acknowledge the computational resources provided by Digital Research Alliance of Canada.

REFERENCES

1. G. J. Brostow, J. Fauqueur, and R. Cipolla, "Semantic object classes in video: A high-definition ground truth database," *Pattern Recognition Letters*, vol. 30, no. 2, pp. 88–97, 2009.
2. A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The KITTI dataset," *International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

3. E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Advances in Neural Information Processing Systems*, vol. 34, pp. 12077–12090, 2021.
4. V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.
5. S. Jégou, M. Drozdzal, D. Vazquez, A. Romero, and Y. Bengio, "The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation," in *IEEE CVPR Workshops*, pp. 11–19, 2017.
6. A. Kendall, V. Badrinarayanan, and R. Cipolla, "Bayesian SegNet: Model uncertainty in deep convolutional encoder-decoder architectures for scene understanding," in *British Machine Vision Conference (BMVC)*, 2016.
7. Y. Tang, L. Wang, and W. Zhao, "Skip-SegFormer: Efficient semantic segmentation for urban driving," in *IEEE International Conference on Intelligent Systems*, 2023.
8. L. Zhao, X. Wei, and J. Chen, "CFF-SegFormer: Lightweight network modeling based on SegFormer," *IEEE Access*, 2023.
9. L. Hoyer, D. Dai, and L. Van Gool, "DAFormer: Improving network architectures and training strategies for domain-adaptive semantic segmentation," in *CVPR*, 2022.
10. B. Cheng, A. Schwing, and A. Kirillov, "Masked-attention mask transformer for universal image segmentation," in *CVPR*, 2022.
11. X. Xu, Y. Li, B. Wu, and W. Yang, "PIDNet: A real-time semantic segmentation network inspired by PID controllers," in *ECCV*, 2022.
12. Y. Zhang, K. Li, K. Chen, *et al.*, "RTFormer: Efficient design for real-time semantic segmentation with transformer," in *CVPR*, 2022.
13. J. Li, T. Xiao, Y. Li, J. Wang, and D. Lin, "FeedFormer: Revisiting transformer decoder for efficient semantic segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.
14. K. Zhao, X. Wang, and Y. Liu, "Fast-SegNet: Fast semantic segmentation network for small objects," *IEEE Transactions on Intelligent Transportation Systems*, 2024.