Proceeding Series of the Brazilian Society of Computational and Applied Mathematics, v. 12, n. 1, 2026.

Trabalho apresentado no XLIV CNMAC, Fundação Getulio Vargas - Rio de Janeiro - RJ, 2025.

Proceeding Series of the Brazilian Society of Computational and Applied Mathematics

CrackFormerSV2: Advanced Pavement Crack Segmentation with Swin Transformer V2 and Dual Attention Mechanisms

Fredy G. Ramírez-Villanueva[1]
FCyT-UNCA, Cnel. Oviedo, Py
José L. Vázquez-Noguera[2], Horacio A. Legal-Ayala[3], Julio C. Mello-Román[4], Pastor E. Pérez-Estigarriba[5]
FP-UNA, San Lorenzo, Py

Pavement crack detection and segmentation is a critical component in the maintenance and monitoring of transportation infrastructure. Although recent advances have been made, significant challenges persist due to the inherent characteristics of pavement cracks.

Recent developments in **Vision Transformers** (ViTs) [2] and the **Swin Transformer** [5] have demonstrated superior performance on numerous computer vision tasks; nevertheless, their straightforward application to crack segmentation is hampered by high computational complexity and a lack of domain-specific inductive biases.

In this work, we introduce **CrackFormerSV2**, a novel *encoder-decoder* architecture expressly designed for pavement crack segmentation. Our framework capitalizes on the capabilities of **Swin Transformer V2** [4] and integrates dual attention mechanisms to overcome the limitations encountered by existing approaches.

CrackFormerSV2 incorporates two complementary attention modules. In the decoder blocks, we integrate the **CBAM** (Convolutional Block Attention Module) [6] to refine feature representations via sequential channel and spatial attention. Additionally, we propose a **Skip Attention** module that augments traditional skip connections. An **ASPP** (Atrous Spatial Pyramid Pooling) [1] module at the bottleneck further consolidates multi-scale contextual information, which is essential for accurately capturing crack patterns of varying widths.

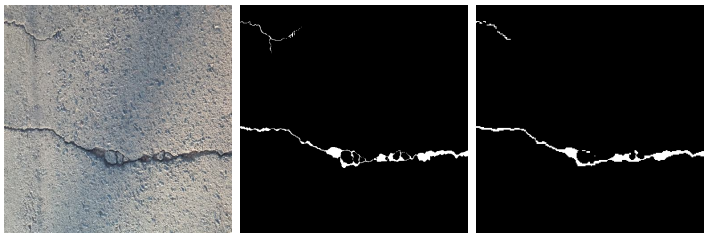
The main innovation of CrackFormerSV2 lies in its encoder architecture, which adopts the Swin Transformer V2. This contrasts with CrackFormer-II, which employs custom transformer encoder modules [3].

To optimize the training process, we employ a scheduling strategy that commences with a learning rate warm-up phase, followed by a cosine annealing schedule. Notably, we assign distinct learning rates to different network componentswith the pre-trained encoder receiving a lower rate (typically one-tenth) relative to the more intensively trained decoder.

We evaluate CrackFormerSV2 on established public benchmarks, such as Crack500 [6], as well as on our proprietary dataset, Fig. 1 shows an example. Compared to a baseline **UNet-ResNet** model which achieved an IoU of 0.466, recall of 0.642, precision of 0.629, and an F1 score of 0.635. CrackFormerSV2 demonstrates projected improvements with IoU values in the range of 0.565 - 0.620 (a 21.2 - 33.0% improvement), recall between 0.725 - 0.780 (a 12.9 - 21.5% improvement), precision between 0.735 - 0.780 (a 16.9 - 24.0% improvement), and F1 scores from 0.730 - 0.780 (a 15.0 - 22.8% enhancement).

[1]framirez@fctunca.edu.py
[2]jlvazquez@pol.una.py
[3]hlegal@pol.una.py
[4]juliomello@pol.una.py
[5]peperez.estigarribia@pol.una.py

© 2026 SBMAC



(a) Image (b) Mask (c) Prediction

Figure 1: Example segmentation prediction. Source: from authors.

With further optimization incorporating gradual unfreezing and more appropriate loss functions, our approach holds significant potential to achieve metric values that are not only competitive with but may well surpass state-of-the-art results.

Acknowledgments

This research (INIC01-302) is co-funded by CONACYT, Paraguay, with the support of FEEI.

References

[1] 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". In: **IEEE Transactions on Pattern Analysis and Machine Intelligence** 40.4 (2018), pp. 834–848. DOI: 10.1109/TPAMI.2017.2699184.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: **arXiv:2010.11929** (2020). DOI: 10.48550/arXiv.2010.11929.

[3] H. Liu, J. Yang, X. Miao, C. Mertz, and H. Kong. "CrackFormer Network for Pavement Crack Segmentation". In: **IEEE Transactions on Intelligent Transportation Systems** 24.9 (2023), pp. 9240–9252. DOI: 10.1109/TITS.2023.3266776.

[4] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, and B. Guo. "Swin Transformer V2: Scaling Up Capacity and Resolution". In: **CoRR** (2021). DOI: 10.48550/arXiv.2111.09883.

[5] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo. "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". In: **arXiv:2103.14030** (2021). DOI: 10.48550/arXiv.2103.14030.

[6] 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)**. Springer, 2018, pp. 3–19. DOI: 10.1007/978-3-030-01234-2_1.

© 2026 SBMAC