Title: Hybrid Vision Transformer and Convolutional Network for Automated Road Damage Detection

Author names: Abu Kausar, Mohidul Islam, Md. Salah Uddin, Abu Shahed Shah Md Nazmul Arefin S. M. Ishtiaque Ahammed Khan Ishti, and Kazi Jahid Hasan

Abstract: The possibility to monitor transportation safety and provide smart infrastructure management depends on the efficient and precise detection of the damage of road surfaces. Nevertheless, traditional convolutional models are prone to the failure to capture complex patterns of context when dealing with a variety of real-world situations. In this paper, we present a new hybrid deep learning architecture combining Convolutional Neural Networks (CNNs) with Vision Transformers (ViTs) with Shifted Patch Tokenization (SPT) and Multi Head Self Attention (MHSA). The model uses CNNs to extract fine-grained local textures and ViTs to extract global spatial features, which makes the framework more effective to capture details of irregularities on the surface (like cracks, potholes, and asphalt deformation). The model development and validation were performed with the help of a domain-specific dataset of 1,600 high-resolution road images, gathered in various districts in Bangladesh. The ViT-ResNet50 hybrid model demonstrated the state-of-the-art results of 99.1% accuracy of classification, surpassing other variants of MobileNetV2-, EfficientNetB0-, and DenseNet121-based models. In addition, Explainable AI (XAI) methods, Grad-CAM, LIME, were added to visualize the areas that define the decisions of models, which improves their model transparency and credibility. These findings validate the importance of employing SPT and MHSA in enhancing both discriminative power and interpretability by a significant margin, representing one large step towards real-time, interpretable, and deployable road damage detection systems in intelligent transportation and smart city ecosystem.

https://authorea-prod.literatumonline.com/doi/full/10.22541/au.176035713.38000317/v1