High-Fidelity Visual Structural Inspection Through Transformers And Learnable Resizers

Kareem ELTOUNY, Seyedomid SAJEDI, Xiao LIANG

Visual inspection is the predominant technique for evaluating the condition of civil infrastructure. Recent advances in unmanned aerial vehicles (UAVs) and artificial intelligence have made visual inspections faster, safer, and more reliable. Camera-equipped UAVs are becoming the new standard in the industry by collecting massive amounts of visual data for human inspectors. Meanwhile, there has been significant research on autonomous visual inspection using deep learning algorithms, including semantic segmentation. While UAVs can capture high-resolution images of buildings' façades, high-resolution segmentation is extremely challenging due to the high computational memory demands. Typically, images are uniformly downsized at the price of losing fine local details. Contrarily, breaking the images into multiple smaller patches can cause a loss of global contextual information. We propose a hybrid strategy that can adapt to different inspection tasks by managing the global and local semantics trade-off. The framework comprises a compound, high-resolution deep learning architecture equipped with an attention-based segmentation model and learnable downsampler-upsampler modules designed for optimal efficiency and information retention. The framework also utilizes vision transformers on a grid of image crops aiming for high precision learning without downsizing. An augmented inference technique is used to boost performance and reduce possible loss of context due to grid cropping. Comprehensive experiments have been performed on 3D physics-based graphics models synthetic environments in the Quake City dataset. The proposed framework is evaluated using several metrics on three segmentation tasks: component type, component damage state, and global damage (crack, rebar, spalling).