**SideRT: A Real-time Pure Transformer Architecture for Single Image Depth Estimation**

Researchers have spent a great deal of time and effort gathering global context since context modelling is essential for determining depth from a single image. For conventional CNN-based architectures, numerous global manipulations are created to get around the locality of convolutions. A better option would be attention mechanisms or transformers that were first developed to capture long-range relationships, but they typically complicate designs and could slow down inference speed. In this study, we offer SideRT, a pure transformer architecture that achieves outstanding real-time predictions. Cross-Scale Attention (CSA) and Multi-Scale Refinement (MSR) modules are created to collaborate in order to effectively fuse features of many sizes in order to better capture global context.

While MSR modules seek to fuse features at corresponding positions, CSA modules concentrate on fusing features with strong semantic similarity. These two modules have a few learnable parameters without convolutions that serve as the foundation for the development of a simple yet powerful model. This architecture has state-of-the-art real-time performance (51.3 frames per second), and it becomes substantially faster with only a modest performance hit by using a smaller backbone Swin-T. (83.1 FPS). Additionally, its performance significantly outperforms the prior state-of-the-art, enhancing the AbsRel metric on KITTI by 6.9% and on NYU by 9.7%. To the best of our knowledge, this is the first study to demonstrate how transformer-based networks may achieve cutting-edge performance in the single image depth estimation sector while operating in real-time. Soon, code will be made accessible.

Two straightforward yet effective approaches are suggested in this research for obtaining improved global context. Using those methods, we construct a simple pure transformer architecture with a few learnable parameters and no convolutions. Our model exhibits strong context modelling capabilities, resulting in cutting-edge results on two difficult datasets. This work shows that a pure transformer design can successfully strike a balance between accuracy and running-time effectiveness. Future research will benefit from these discoveries, which will motivate scientists to focus more on creating real-time transformer designs for useful applications.