EquiSegformer in dataset shift due to panoramic images

Sanskar Singh [1] Gautam Gupta [1]

[1]IIIT Naya Raipur

Problem Statement

- **Research Challenge**: Developing robust panoramic segmentation models faces a significant hurdle due to dataset shift.
- **Limitation**: Pixelwise annotations designed for narrow-angle imagery are insufficient for effective model training.
- **Hurdles**: Inherent distortions and unique feature distributions in 360-degree panoramas pose substantial challenges.
- **Performance Gap**: Transferring knowledge from annotation-rich pinhole camera images towards panoramic tasks results in a noticeable performance deficit.

Proposed Solution

- An initial **equi-rectangular** based convolution, EquiConv, is applied for extracting distortion-aware features in panoramic images.
- Extracted feature maps are fed through an encoder-decoder-based **Segformer**[4] architecture to finally attain segmented masks in the images.
- **EquiSegformer** based design demonstrates a decent efficacy in capturing geometrically aware features from the generated maps while being computationally inexpensive and simpler.

Datasets

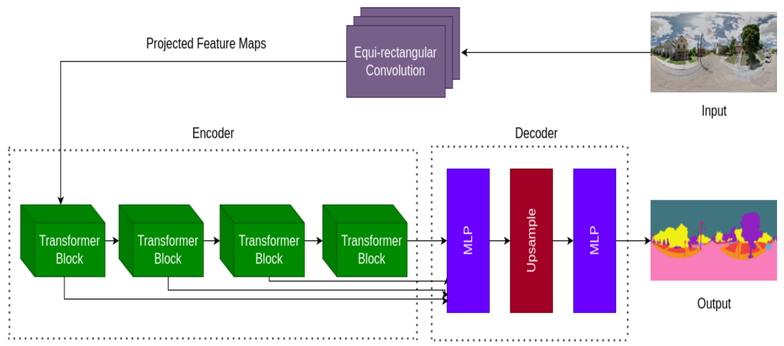
We utilized 3 datasets to validate the proposed prototype. Following are the datasets along with their brief descriptions:

- **CVRG-Pano[3] :** Semantic annotation of outdoor panoramic images, akin to the CityScrapes dataset. It includes 600 images with pixel-level annotations, featuring 20 semantic classes categorized into 7 groups.
- **Extended SUN360[1] :** It offers a comprehensive analysis of primary objects within an indoor scene using a single 360-degree image in equirectangular projection. The dataset consists of 666 indoor panoramas sourced from Zhang et al.'s SUN360 dataset, covering 14 distinct classes of objects associated with main indoor scenes.
- **F-360 SOD[2] :** Contains 500 high-resolution equirectangular images for a 360-degree image-based Salient Object Detection(SOD). It was collected through the extraction of frames from 5 mainstream 360-degree videos.

Model Architecture

EquiSegformer majorly consists of 2 parts that is :

- EquiConv : It applies deformable convolution and projects the image into a spherical domain. The pixels are then inverse projected using inverse gnomic projection of 3D kernel coordinates. Thus, EquiConv applied learned offsets to kernel locations rather than a grid-like structure.
- Segformer : It leverages transformers for semantic segmentation by introducing a hierarchical design. At its core, the model consists of multiple stages, each incorporating a Transformer Encoder block. The final segmentation map is generated by the decoder, which involves a combination of upsampling and convolutional layers



Results

We have utilized Mean Intersection Over Union(Mean IoU) for the evaluation of our proposed prototype for better comparison with previous results in the semantic segmentation domain. IoU can be given as :

$$IoU = \frac{Areaofintersection}{Areaofunion} \quad (1)$$

The following table depicts the achieved results on the experimented datasets:

Datasets Utilized	Previous	EquiSegformer
CVRG-Pano	0.673	**0.675**
Extended SUN360	0.56	**0.94**
F-360 SOD	0.62	**0.715**

Table 1. Calculated Evaluation Metrics in terms of Mean IoU

Conclusion and Future Works

We present the **EquiSegformer** architecture, a lightweight solution for semantic segmentation of panoramic images that seamlessly integrates into existing frameworks with minimal adjustments to transformer blocks. Our approach employs a Segformer-based methodology to tackle dataset shifts effectively. It achieves a **SOTA** accuracy on the defined datasets. While demonstrating impressive performance on smaller datasets, there is a recognized need for substantial improvement in handling larger dataset containing more varied samples.

References

[1] J. Guerrero-Viu, C. Fernandez-Labrador, C. Demonceaux, and J. J. Guerrero. What's in my room? object recognition on indoor panoramic images. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 567–573, 2020.

[2] J. Li, J. Su, C. Xia, and Y. Tian. Distortion-adaptive salient object detection in 360° omnidirectional images, 2019.

[3] S. Orhan and Y. Bastanlar. Semantic segmentation of outdoor panoramic images. *Signal, Image and Video Processing*, 16, 04 2022.

[4] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo. Segformer: Simple and efficient design for semantic segmentation with transformers, 2021.