In recent years, Virtual Reality (VR) has gained widespread popularity, providing users with highly immersive experiences and allowing them to feel as if they were in a virtual world distinct from reality. Omnidirectional images and videos have become one of the most popular content types for VR, thanks to the availability of user-friendly and low-cost 360-degree cameras. This content allows users to be placed at the center of a sphere and freely explore the environment in any direction by simply moving their heads. Although 360-degree media allow increased user immersion, their processing and transmission still entails many challenges. One major issue is the high memory and bandwidth requirements with respect to standard 2D content. For example, streaming a 4K 2D video requires approximately 25 Mb/s, while delivering a 4K resolution for each eye to provide a full 360-degree viewing experience requires around 400 Mb/s. One effective way to address the transmission challenges of 360-degree videos is to implement a user-centered streaming paradigm. To this aim, human attention mechanisms have been studied to design saliency estimation methods. These algorithms compute 2D probability maps which highlight the regions inside a 360-degree scene most likely to draw users' attention. These maps can then be used to transmit the salient regions at higher quality while encoding at lower quality (or discarding) less relevant areas.

We introduced **Sphere-GAN**, a model designed to estimate saliency maps for 360° videos by taking as input the current frame in equirectangular form at time t, along with the ground-truth saliency map from $t - 5$. The architecture follows a GAN framework, where the generator adopts a U-Net structure and the discriminator consists of four 2D convolutional layers. To effectively address the geometric distortions near the poles inherent in the equirectangular projection, we replaced standard 2D convolutions in the generator with spherical convolutions. Unlike standard 2D convolutions, spherical convolutions operate directly on the tangent plane of the sphere. In this process, the convolutional kernel is applied locally to tangent planes across the spherical surface, allowing each convolution operation to adapt naturally to the local curvature of the sphere. Consequently, when 360° frames are projected onto the equirectangular domain, the convolutional kernels are subject to the same geometric distortions as the frames, thereby addressing distortion issues and reducing discontinuities and oversampling near the poles. Through this design, the model achieves more accurate and perceptually consistent saliency estimation for 360° videos.

Using the Sport360 dataset, which consists of 104 360° videos watched by 20 users at least for every video, Sphere-GAN achieved superior performance compared to state-of-the-art baselines, evaluated with metrics such as Pearson Correlation Coefficient (CC), Kullback-Leibler Divergence (KL), Normalized Scanpath Saliency (NSS), and Area Under the Curve–JUDD (AUC_JUDD). Our approach outperforms the best state-of-the-art method (SphereU-Net) by achieving an 8.5% higher in CC, a 19.8% improvement in NSS, an 81.7% reduction in KL, and a 17% gain in AUC_JUDD.