Exploring Spherical Autoencoder for Spherical Video Content Processing

Jin Zhou, Na Li, Shuochao Yao, Yao Liu, Songqing Chen

George Mason University, Rutgers University

1.Abstract

3D spherical content is increasingly presented in various applications (e.g., AR/MR/VR) for better users' immersiveness experience, yet today processing such spherical 3D content still mainly relies on the traditional 2D approaches after projection, leading to the distortion and/or loss of critical information. We propose a novel approach called Spherical Autoencoder (SAE) for spherical video processing. SAE represents the 360-degree video content as a spherical object and employs encoding and decoding on the 360-degree video directly. To support the adoption of SAE on pervasive mobile devices that often have resource constraints, we further propose two optimizations on top of SAE - a SAE scheme with the partial view support that can utilize such FoV prediction and Compressive Sensing. Our extensive experiments show that directly incorporating and processing spherical signals is promising, and it outperforms the traditional approaches by a large margin.

3. Optimization

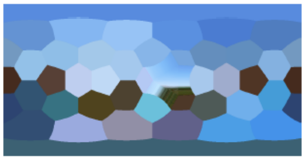
- **Partial Mesh Spherical Autoencoder**

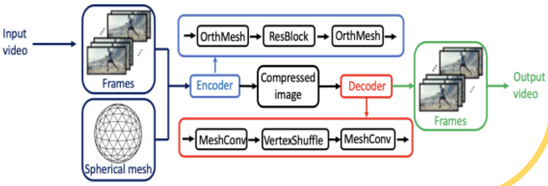
p-SAE is designed to utilize the field-of-view (FoV) prediction. To support the utilization of FoV, we use the partial icosahedral mesh. The partial icosahedral mesh is created by selecting one triangular face from the full Level-1 mesh (that is, only 1 out of the

Figure is the example image after using partial mesh 80 faces) and only refine triangles within this face. As a result, the refined face is about 1/80 of the sphere and contains roughly 1/80 of the vertices in a full mesh.

- **Compressive Sensing in SAE**

To address the drawbacks, compressive SAE (c-SAE) links the proposed SAE (spherical encoder-decoder structure). with compressive sensing theory. According to the recent deep compressive offloading theory, we should impose the Restricted Isometry Property (RIP) (with orthogonal regularization) and Lipschitz continuity (with spectral normalization) on the spherical encoder and decoder, respectively, to provide recovery guarantees for the data encoding-decoding process based on compressive sensing theory.





2.SAE Structure

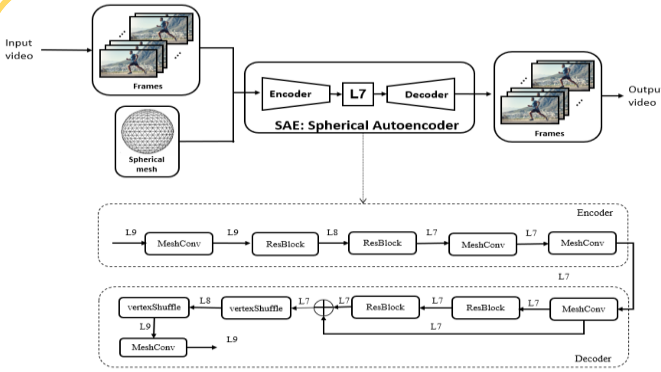


Figure shows the proposed spherical autoencoder (SAE) architecture.

To process an input video, we first load each frame into the icosahedral mesh. The encoder of SAE loads the RGB values of pixels on the spherical video frame as values of vertices on the icosahedral mesh, It then goes through a MeshConv layer with batch normalization and ReLU function. The output then goes through two ResBlocks to both coarsen the mesh (i.e., coarsen the mesh from Level-9 mesh to Level-8 and Level-7 meshes, respectively.) and increase the channel dimension.

Finally, we use another MeshConv layer to change the channel dimension to 3. The decoder passes its input tensor through one MeshConv layer and two ResBlocks to increase the channel dimensions. It then passes the output through two VertexShuffle operations to increase the number of vertices. A final MeshConv layer in the decoder produces a 3 × Nv,9 tensor, in the same dimension as the original input to the SAE encoder.

4.Experiments Result

PSNR result of CAE, SAE, p-SAE, c-SAE, c-p-SAE with different compression rates

Model	CAE	SAE				p-SAE			
Model Compression Ratio	none	none	26.01%	41.54%	58.30%	none	36.03%	45.97%	55.11%
Indoor	24.7384	39.2887	41.2183	39.0652	38.3835	37.8229	37.7869	36.4909	35.9083
City	17.1524	39.7354	38.9868	38.5078	37.6546	33.4587	33.6031	32.7565	31.8372
Roller-Coaster	17.8414	34.1936	32.8375	32.0299	31.5185	32.5046	32.1746	31.5830	30.3547
Football	20.6538	36.2050	36.3093	36.2165	36.1787	34.4323	33.9375	33.7339	33.2091
Model	co-CAE	c-SAE				c-p-SAE			
Model Compression Ratio	37.21%	none	26.01%	41.54%	58.30%	none	36.03%	45.97%	55.11%
Indoor	20.1124	40.5602	41.3595	39.7724	38.7182	38.6803	38.5517	37.6399	37.5349
City	15.0681	40.3514	39.4804	39.1991	38.9915	34.5360	36.4247	33.6042	33.0823
Roller-Coaster	13.5168	35.5432	33.8656	33.4480	32.6312	33.2020	32.2679	32.1994	31.0052
Football	14.8910	35.5992	36.8822	36.4333	36.0842	36.6811	36.8463	36.4243	36.1195

Table reports the PSNR results of our SAE model and its variants. In these models, the Level-9 mesh is used. In this table, CAE represents the results from the traditional 2D convolutional autoencoder, and co-CAE represents the compressed CAE. SAE (Spherical AutoEncoder) uses the full sphere mesh. p-SAE uses partial mesh in SAE instead of full mesh. c-SAE applies compressive sensing in SAE, and c-p-SAE applies compressive sensing in p-SAE. These models are all compressed with different compression rates as indicated in the table (shown as "model compression ratio").

5.Conclusion

- ➢ Compared to the traditional 2D video, spherical video content not only demands more bandwidth to transmit, but also more efficient techniques for content processing. In this work, we explored a new approach to effectively process spherical content.
- ➢ Compared to the traditional approach where a spherical frame is mapped to a 2D space, we have investigated processing the spherical content directly using a spherical autoencoder (SAE).
- ➢ Motivated by the fact that mobile devices are widely used for video accesses, we have further proposed two optimizations to make SAE better fit for resource constrained mobile devices while maintaining the video quality.