A New Dataset and Transformer for Stereoscopic Video Super-Resolution

Hassan Imani[1], Md Baharul Islam[1,2], Lai-Kuan Wong[3]

[1]Bahcesehir University, [2]American University of Malta, [3]Multimedia University

Summary

Background: Stereo video super-resolution (SVSR) aims to enhance the spatial resolution of the low-resolution video by reconstructing the high-resolution video.

Motivation: Transformer-based models for vision tasks such as Vision Transformers (ViT) [1] which divide a video frame into small patches and extract the global relationships among the token embeddings, cannot be directly applied for SVSR, in which the local and texture information is essential. Furthermore, temporal information and consistency, which are equally crucial in the SVSR task, cannot be solved by ViT.

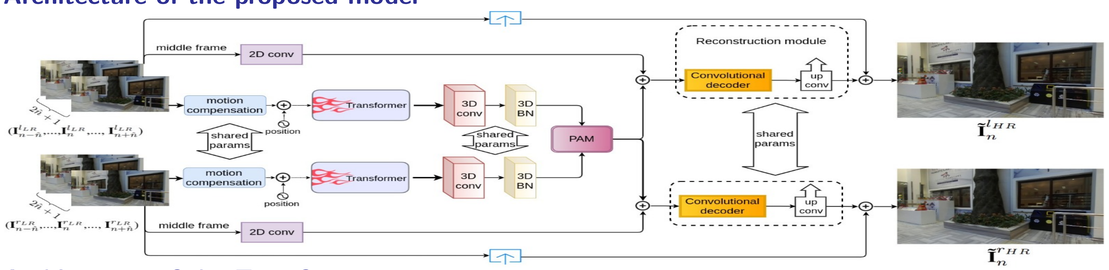
Contribution: We propose a novel Transformer-based model that can integrate the spatio-temporal information from the stereo views while maintaining both stereo- and temporal consistency. Furthermore, a new dataset, namely *SVSR-Set*, is collected for the SVSR task. It contains 71 high-resolution stereo videos in different indoor and outdoor settings, and is the largest dataset for the SVSR task.

SVSR-Set Dataset

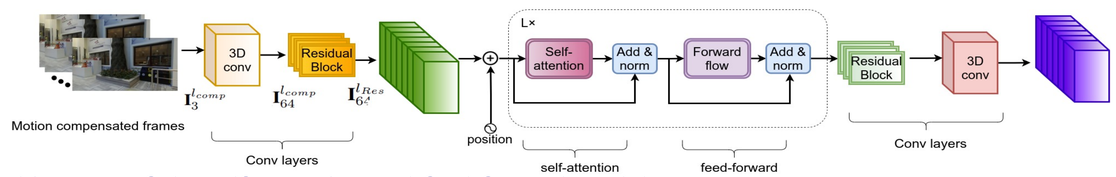
Stimuli Type	# videos	Light	Motion	Setting
people,tree	5	day	high	Outdoor
people,tree,car,motor	12	day	high	Outdoor
people,tree,car,motor	4	night	high	Outdoor
people,train	2	day	high	Outdoor
people,ship,water	9	day	low	Outdoor
bird,dog,grass	4	day	low	Outdoor
grass,motor	3	day	high	Outdoor
water,bird	8	day	high	Outdoor
people	5	night	low	Outdoor
toy	6	day	high	Indoor
game	2	day	high	Indoor
flower	4	day	high	Outdoor
people	7	day	low	Indoor

Method

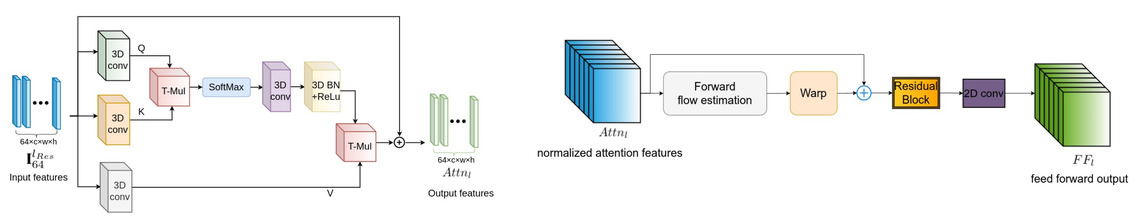
Architecture of the proposed model



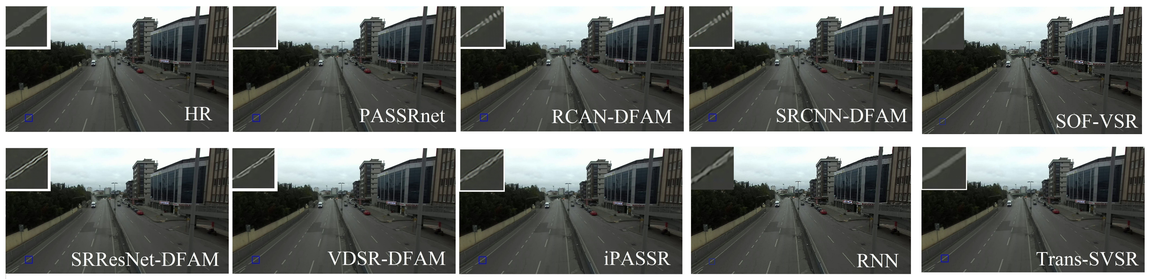
Architecture of the Transformer



Architecture of the self-attention and feed-forward modules



Experiments



Performance comparison of our proposed Trans-SVSR and state-of-the-art Stereo ISR methods:

Dataset		SVSR-Set		NAMA3D		LFO3D	
Stereo ISR methods	N.Par	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
PASSRnet [2]	1.41M	28.9723	0.8957	25.8708	0.8335	22.4558	0.7047
SRRes+SAM [3]	1.73M	27.2863	0.8681	24.2226	0.7978	19.4583	0.6736
SRResNet-DFAM [4]	2.89M	27.5388	0.8905	24.8192	0.8212	21.0419	0.7389
SRCNN-DFAM [4]	0.73M	27.7202	0.9008	24.5390	0.8344	21.2752	0.7355
VDSR-DFAM [4]	2.68M	28.4919	0.8949	25.2385	0.8401	22.1496	0.7435
RCAN-DFAM [4]	16.9M	29.0158	0.9013	25.9539	0.8442	22.5685	0.7427
iPASSR [5]	1.42M	28.1980	0.8913	24.7818	0.8195	21.4525	0.6865
2D-VSR methods							
SOF-VSR [6]	2.08M	28.7208	0.9002	24.573	0.8401	22.5642	0.7414
RRN [7]	14.40M	29.1454	0.9069	24.6100	0.8418	22.7596	0.7483
SVSR methods							
Trans-SVSR	27.29M	**31.9766**	**0.9293**	**28.8424**	**0.8674**	**25.5871**	**0.7642**

Improvement over Various Data Augmentation Techniques

Model type	N.Par	PSNR	SSIM
Trans-SVSR-WMF	27.29M	28.9715	0.8619
Trans-SVSR-WOT	23.47M	29.8914	0.8959
Trans-SVSR-WOP	27.20M	30.6348	0.9068
Trans-SVSR-WOR	9.48M	30.3605	0.9031
Trans-SVSR-WSF	27.29M	30.3129	0.8989
Trans-SVSR	27.29M	**31.9766**	**0.9293**

References

[1] Dosovitskiy *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. arXiv 2020.
[2] Wang *et al.* Learning parallax attention for stereo image super-resolution. In CVPR 2019.
[3] Ying *et al.* A stereo attention module for stereo image super-resolution. IEEE SPL 2020.
[4] Dan *et al.* A disparity feature alignment module for stereo image superresolution. IEEE SPL 2021.
[5] Wang *et al.* Symmetric parallax attention for stereo image super-resolution. In CVPR 2021.
[6] Wang *et al.* Deep video super-resolution using hr optical flow estimation. IEEE TIP 2020.
[7] Isobe *et al.* Revisiting temporal modeling for video super-resolution. arXiv 2020.

Funded by: TÜBİTAK