Light Field Image Compression Using Multi-Branch Spatial Transformer Networks Based View Synthesis

Jin Wang[#], Qianwen Wang[#], Ruiqin Xiong[+], Qing Zhu[#], Baocai Yin[*]

*(# Beijing University of Technology, + Peking University, * Dalian University of Technology)*

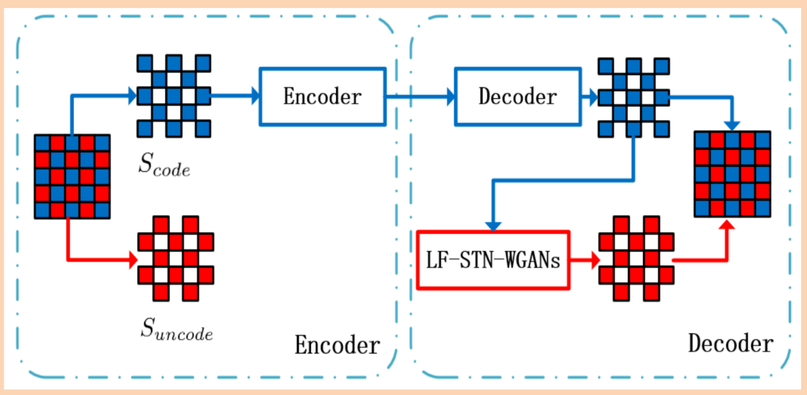
Abstract

In this paper, we propose a novel light field image compression scheme using multi-branch spatial transformer networks based view synthesis. Firstly, a sparse subset of views are selected and are rearranged into a pseudo sequence to be encoded by an video codec at encoder. Then the other unselected views are synthesized based on the similarity between neighboring views with our proposed method at decoder. A multi-branch spatial transformer networks (MSTN) is designed to adaptively learn the affine transformations between the neighboring views, which are used to warp the input views to generate accurate approximation of the target views. Moreover, to better obtain the final view by the generated approximation views, the Wasserstein generative adversarial networks(WGAN) is applied with the improved training.

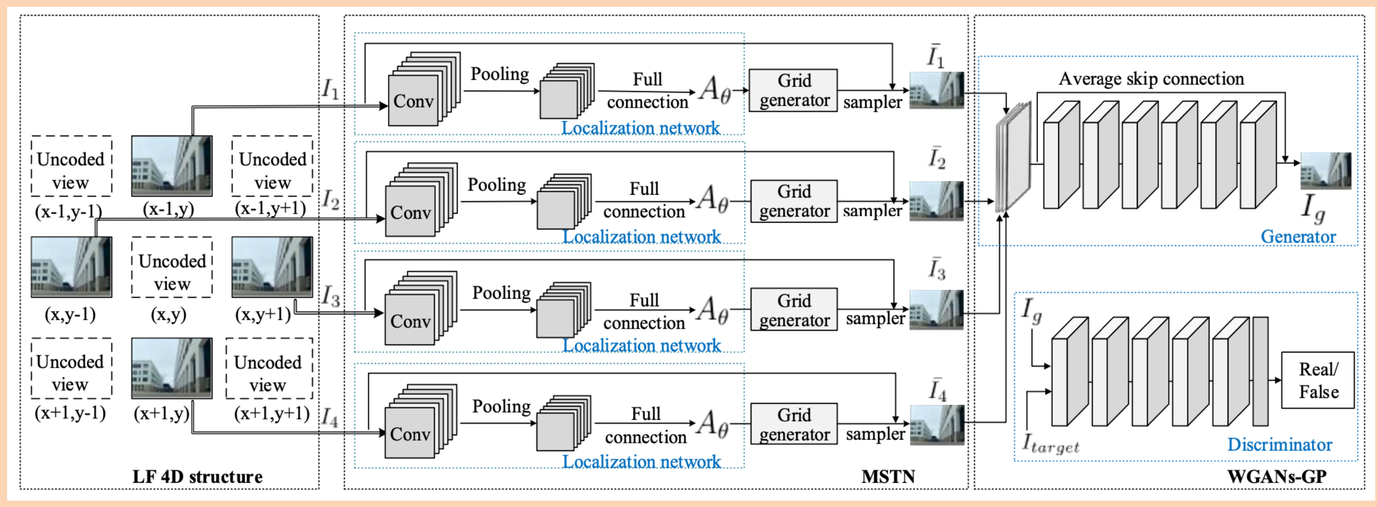
Proposed Scheme

A multi-branch spatial transformer networks (MSTN) is designed to warp the input views to generate the approximation of the unencoded view by affine transformation. MSTN utilizes *p* surrounding (*p* equals the number of branches) views as inputs for the multi-branches, each of which shares the same network structure design but with different weights and bias. The warped views can be represented as:

$$(\bar{I}_1, \bar{I}_2, ..., \bar{I}_p) = F_{MSTN}(I_1, I_2, ..., I_p)$$



After transforming the neighboring views of the unencoded view via MSTN, Wasserstein GAN(WGAN) is further utilized to learn the relationship between the transformed image of *p*-branches and the unencoded view. The unencoded view is estimated from a set of input features including all the transformed images. WGAN tends to directly learn the mapping between such two spaces with the objective of minimizing L2 distance between the unencoded view and the output of WGAN. Both adversarial loss and mean square loss are adopted in our final model to guarantee the high quality signal restoration as well as perceptual quality of our scheme.

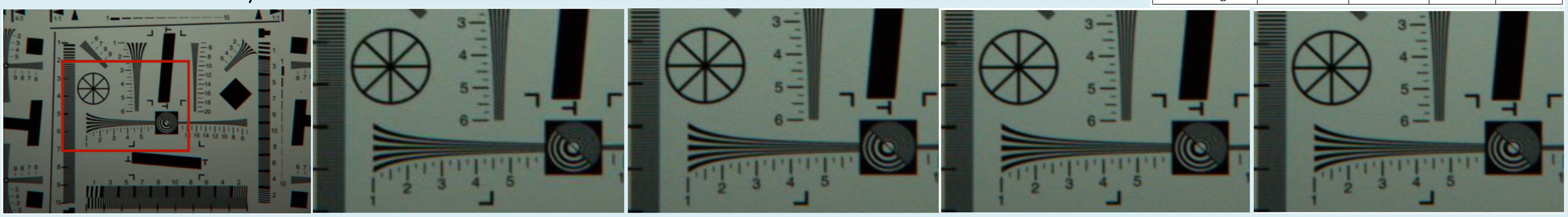


Experimental Results

The EPFL LF image dataset is chosen for training and evaluation. Specifically, twelve LF images from EPFL dataset are selected for testing. For the LF image pre-processing, the lenslet images in these datasets are firstly decomposed into 4-D structure to obtain views. Only the internal 9×9 views are adopted because of the significant decomposed distortion of the other views. The bit depth of these views is 8 bits and the color space is YUV420. To accelerate the convergence of model during the training process, patches of size 32×32 with a stride of 31 pixels are extracted from the full images. All of the image values are normalized to the unit interval [0, 1]. Mini-batches of size 64 is used to get the best trade-off between speed and convergence. Our networks' weights are randomly initialized by the random method.

HM-LDB: HEVC low delay **HM-RA**: HEVC random access **LA-41**: Zhao *et al*. in *ICIP* 2018 **DL-41**: Zhao *et al*. in *ICME* 2018

LF-images	HM-LDB	HM-RA	LA-41	DL-41
*I*01	0.45	0.39	-0.15	0.07
*I*02	0.18	0.21	0.03	0.12
*I*03	0.11	0.17	-0.15	0.07
*I*04	0.65	0.53	0.01	0.14
*I*05	0.42	0.80	0.27	0.16
*I*06	0.27	0.41	0.15	0.10
*I*07	0.40	0.41	-0.03	0.06
*I*08	0.44	0.50	0.24	0.76
*I*09	0.22	0.28	-0.10	0.11
*I*10	0.52	0.79	0.27	0.05
*I*11	0.40	0.60	0.02	0.06
*I*12	0.45	0.32	0.05	0.18
Average	0.38	0.42	0.06	0.16



Test image *ISO* | HM-LDB (34.70dB) | LA-41 (34.64dB) | DL-41 (34.44dB) | Proposed (34.93dB)

Test image *Bikes* | HM-LDB (35.63dB) | LA-41 (36.30dB) | DL-41 (36.44dB) | Proposed (36.68dB)