**Production of DGIST campus 3D Model: Utilizing Data Lightening in SOTA 3D Reconstruction**

Kong Yeong Jae

School of Undergraduate Studies, DGIST

**Abstract**

This research explores the utilization of state-of-the-art 3D reconstruction techniques, specifically focusing on Local Light Field Fusion (LLFF), to create a 3D model of the DGIST campus. Recognizing the challenge of reducing input data for efficient 3D reconstruction, our team introduces a novel approach involving data lightening through frame interpolation. The method incorporates RRIN, a lightweight frame interpolation model, to reduce the dataset size while addressing the challenge of viewpoint differences between frames. To further enhance data lightening, Extreme View Synthesis is introduced, aiming to replace frame-interpolated images. However, encountering challenges in depth-map estimation, we sought to address the issue through the integration of DPSNet to enhance the quality of the depth map. This process aimed to overcome limitations associated with directly applying Extreme View Synthesis to LLFF, particularly due to the complexities involved in estimating camera parameters for distant views. Our study represents a valuable exploration into advancing depth-map quality and extending the applicability of frame interpolation for efficient 3D reconstruction.

**Production of DGIST campus 3D Model: Utilizing Data Lightening in SOTA 3D Reconstruction**

Reconstructing the real world has long been a goal in the field of computer vision. As part of this, research is being actively conducted to expand 2D images to 3D domains, such as 3D reconstruction based on Deep Learning. This field holds significance due to its extensive application across various domains, including medicine, robotics, the film industry, urban planning, and virtual environments (Samavati & Soryani, 2023). Notably, the field of view-synthesis has become active with the advent of Neural Radiance Fields (NeRF), which can do view-synthesis in the unconstrained image set (Mildenhall et al., 2020, p.1). NeRF possesses the remarkable ability to synthesize views from an infinite dataset of images. Of significance here is the introduction of positional encoding technology, which has allowed for comprehensive view synthesis in all directions (Mildenhall et al., 2020, p.7). This innovation primarily relies on Multi-Layer Perceptrons (Murtagh et al., 1991) instead of the traditionally used Convolutional Neural Networks (O`Shea & Nash, 2015). This method allowed reconstruction of places that could not be seen only with sparsely sampled views (Mildenhall et al., 2020, p.2).

Local Light Field Fusion (LLFF), the previous state-of-the-art of 3D reconstruction fields (Mildenhall et al., 2019), also showed performance comparable to NeRF, but the conditions were a little complex and difficult to use in practice. In addition, NeRF combines differentiable rendering technology to produce higher quality reconstruction than LLFF. However, NeRF also has a problem that high-quality output can be derived only when there are at least 20 different view-direction images. As a result, models such as GIRAFFE, which merged NeRF into the generative model have also emerged (Niemeyer & Geiger, 2021, p.2), but these generative models can be difficult to directly control the view-synthesis scene. In order to control the scene, there is a method of remembering the application and shape code of the scene to view-synthesize through Gan inversion (Xia et al., 2021, p.1) and then manipulating the cam-pose while continuing to refer to it, but this also has the disadvantage of consuming too much inference cost and poor novelty in the optimization process of Gan inversion.

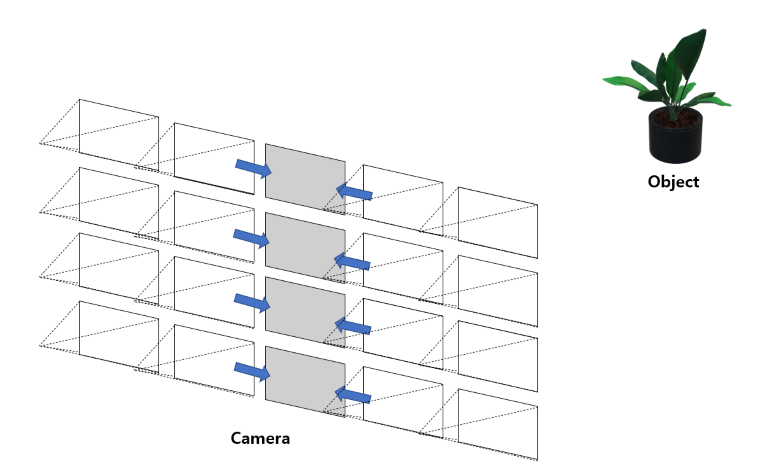
In this work, our team identified a significant gap in the field concerning the reduction of input data required for 3D reconstruction, a persistent challenge that has constrained the efficiency and applicability of reconstructive technologies across various domains. We conducted 3D reconstruction for DGIST objects, and further conducted research to reduce the number of data sets required by the existing framework. Clearly, a column was removed from the existing data set to lighten the data. After that, we run frame interpolation with images on both sides of the removed column, putting it into LLFF input and proceeding with a 3D reconstruction with less data.

**Methodology**

Among the 3D reconstruction works, 3D reconstruction with 4,000 times fewer images than previous studies is enabled by Local Light Field Fusion (LLFF) (Mildenhall et al., 2019, p.1). The targeted objects can be reconstructed in 3D by collecting images in a total of 20 grids, composed of 4x5 forms. Consequently, it was hypothesized that recent technologies could be integrated to enable 3D reconstruction with datasets lighter than these 20. Moreover, frame interpolation was involved in the methods utilized to lighten the input data for LLFF. Data lightening was achieved by implementing frame interpolation and integrating images on both sides of the column, extracted from the existing dataset. Consequently, 3D reconstruction was anticipated to be enabled with merely 16 pieces of data. The overview of the model is indicated in Figure 1.

**Figure 1.**

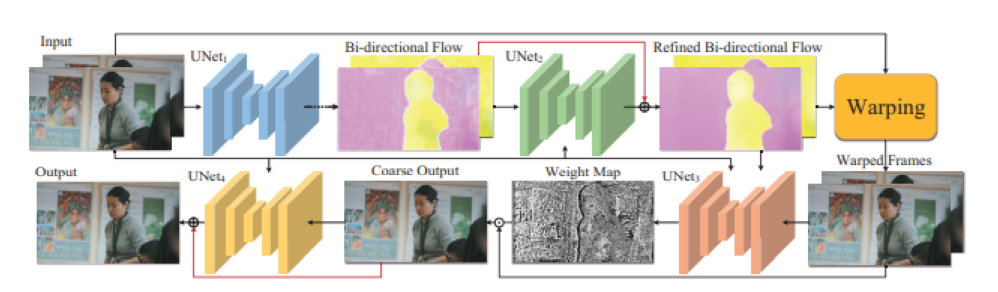
*Data Lightweight Idea Overview*



Among several frame interpolation models, RRIN (Li et al., 2020) was selected for frame interpolation due to its low complexity and strong performance. RRIN has improved performance while reducing the complexity of training by applying reactive learning to the place where optical flow and warped frames are refined (Li et al., 2020, p.2). Additionally, a new adaptive weight map was introduced to render the interpolated frame more visually natural, incorporating the light change and occlusion information of the frame (Li et al., 2020, p.3). Moreover, RRIN significantly reduces the complexity and volume of the model, with all sub-modules being composed solely of shallow UNets (Li et al., 2020, p.2). Despite the model's lightweight nature, quantitative superiority over other works was demonstrated. The model pipeline is depicted in Figure 2.

**Figure 2.**

*Overview of the proposed RRIN*



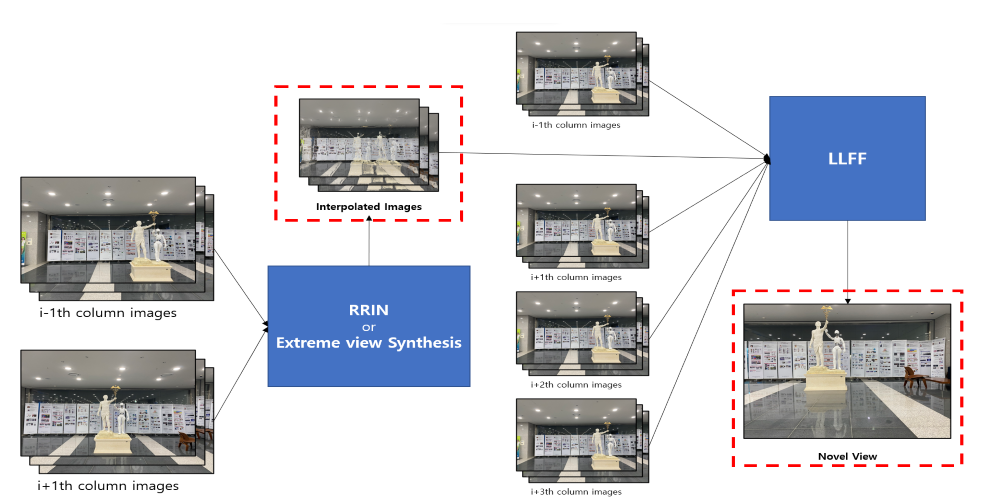
Li, H. (2020). Video Frame Interpolation Via Residue Refinement, *ICASSP 2020,* 2.

<https://doi.org/10.1109/ICASSP40776.2020.9053987>

Adding heavy and complex pipelines to reduce the size of existing LLFF datasets was considered unnecessary. Therefore, RRIN, known for its compact and efficient interpolation was used. However, data lightening performed using our frame interpolation method was also unsatisfactory due to poor quality of the replaced images, attributed to the substantial viewpoint difference between frames. To overcome this, data weight reduction was additionally conducted through the Extreme View Synthesis model (Choi et al., 2019). Extreme View Synthesis was performed Novel View Synthesis with a minimal number of different viewpoint images. This was used to replace the existing frame interpolated images. The entire pipeline is depicted in Figure 3.

**Figure 3.**

*Pipeline of Data Lightening*



**Results**

Figure 4 shows the result of 3D reconstruction based on normalized 4x5 images. Although black noise occurs at the edge of some frame, view synthesis themselves have very few artifacts and smooth good results. Based on the pre-trained model, this experiment was conducted in a computing environment of Tesla P100-PCIE-16GB (GPU) Intel(R) Xeon(R) CPU @ 2.3 GHz (CPU). The test time was 3-4 hours per dataset.

**Figure 4.**

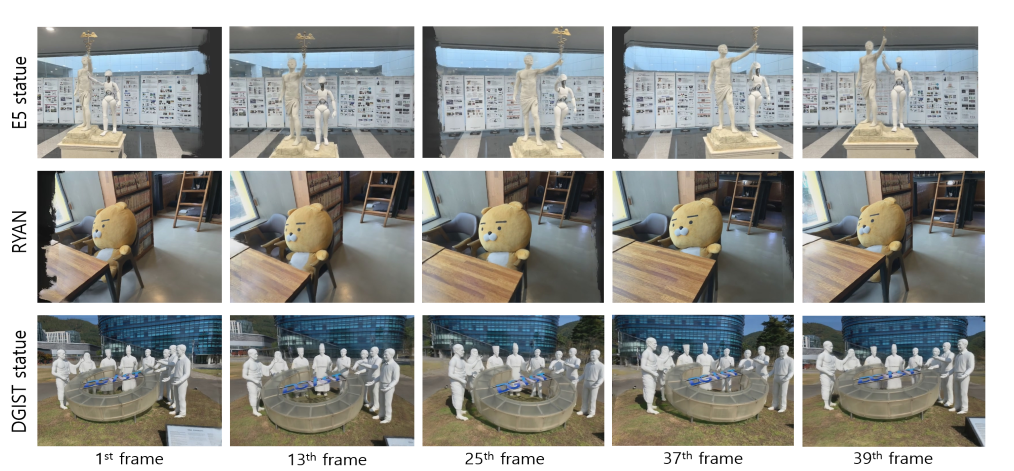
*DGIST object LLFF results* **

Figure 5 shows the results of 3D reconstruction based on data sets replaced by interpolated images for images excluded from existing data sets. In order to accurately compare the effects of Frame interpolation, the results of 3D reconstruction with 4x3 and 4x4 datasets without interpolated images were also compared.

**Figure 5.**

*LLFF Performance Comparison by Interpolated Dataset*

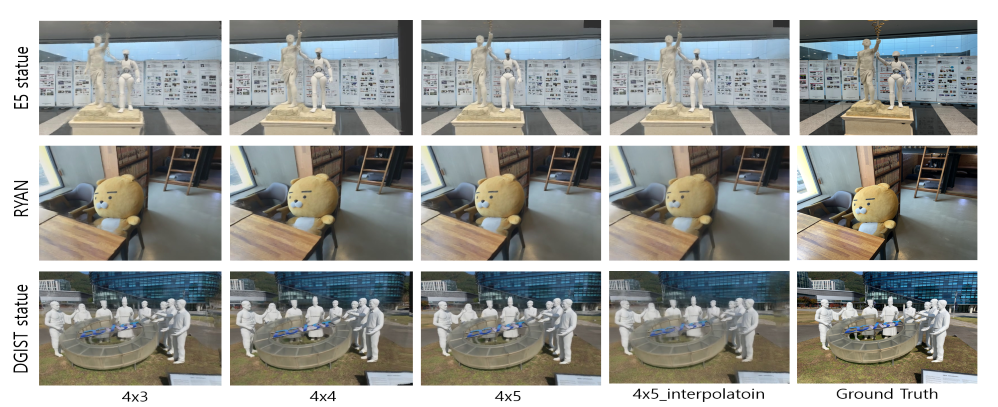
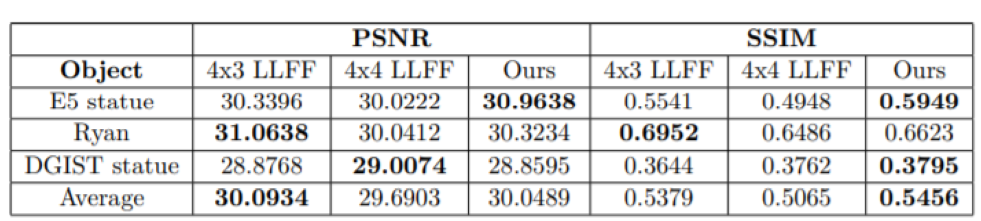


Figure 6 is a quantitative evaluation of the 3D reconstruction data derived above. This quantitative evaluation was conducted based on the following two assumptions. The first is that because the new View was synthesized, it is difficult to take Ground Truth, which exactly matches the View direction and camera pose. Second, frame interpolation is inferior to conventional LLFF because there is a limitation that smooth interpolation is not possible for our dataset targets. Therefore, we quantitatively evaluated the results of our model using the LLFF results targeting the existing 4x5 data as GT. "Ours" in Figure 6 is based on 16 data that removed the data of the second column from the existing data set of 4x5 and 4 data generated by frame interpolation of the data of the first and third columns, and a total of 20 data. The evaluation metrics we used are PSNR and SSIM. PSNR evaluates loss information on the image quality of the generated image, and the lower the loss (the better the image quality), the higher the value. SSIM evaluates the difference between ground-truth and generated images in three aspects: Luminance, Contrast, and Structure. For PSNR, our work tended to be lower than the 4x3 dataset, but higher than the 4x4 dataset. SSIM tended to be higher than other works.

**Figure 6.**

*Quantitative assessment of LLFF results based on dataset*



We tried to apply extreme view synthesis as an alternative model for lightening to solve the problem of many artifacts occurring in frame interpolated images. Figure 7 is the extreme view synthesis result, but not all areas were synthesized except for the background. Extreme View Synthesis used a pre-trained model, tested with Tesla V100's computing specifications, and tested for about 15 minutes per dataset.

**Figure 7.**

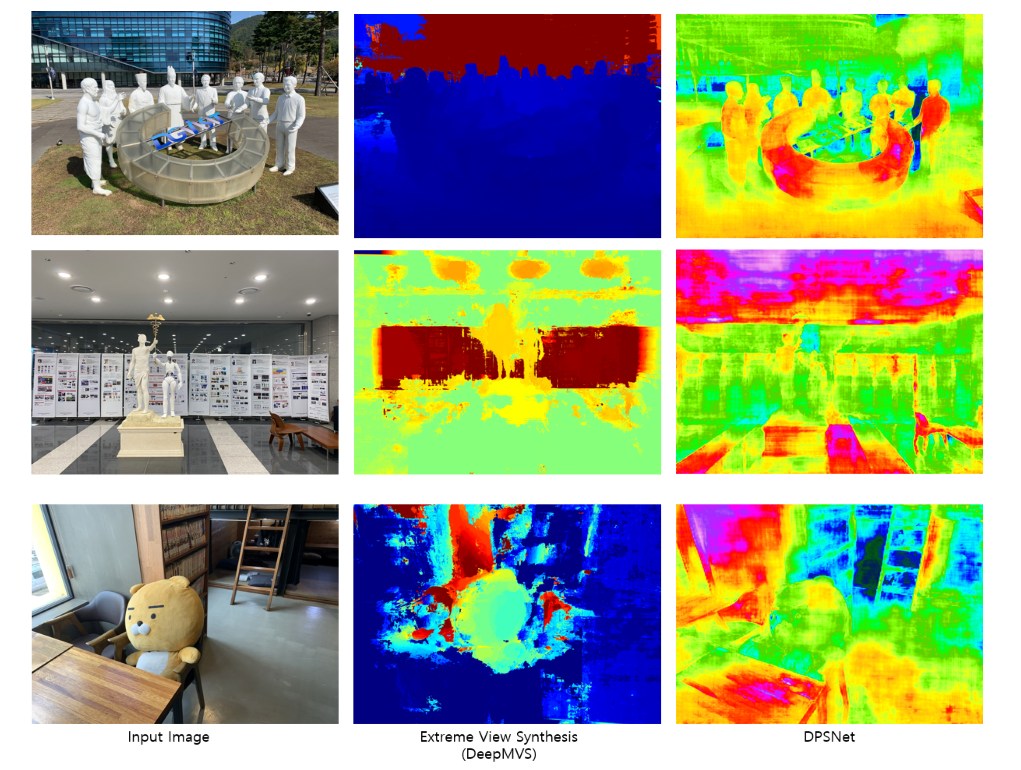
*Two different frames in the results of Extreme View Synthesis*



Therefore, it could not be used as an alternative for data weight reduction. When a depth map was drawn to analyze the cause of these results, it was determined that there was a problem in the depth classification stage of this paper. Therefore, we decided to extract the depth map through the DPSNet model. The comparison results are shown in Figure 8.

**Figure 8.**

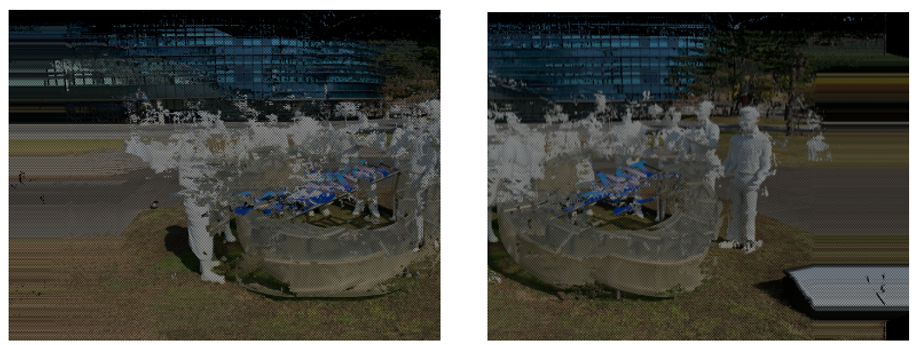
*Comparison of Depth Map in Extreme View Synthesis and DPSNet*



We can see that DPSNet shows significantly better depth compared to Extreme View Synthesis. Therefore, we proceeded with Extreme View Syntheis by replacing the depth-map with the output depth of DPSNet, and Figure 9 shows the results. DPSNet also used a pre-trained model, and was tested with a computing stack of Tesla P100-PCIE-16GB (GPU) Intel(R) Xeon(R) CPU @ 2.3GHz (CPU), and the test time was about 30 minutes per dataset.

**Figure 9.**

*Two different frames in the results of Extreme View Synthesis by DPSNet Depth map*



Although we derived a relatively better Depth map with DPSNet, the results of Extreme View Synthesis were insufficient for use as input for LLFF. In the case of existing Extreme View Synthesis, almost all parts were not synthesized, but after improving the Depth map, synthesis was made in many areas. This proved that the first attempt was a problem with the estimation of Depth. Nevertheless, the poor results of Extreme View Synthesis, which improved depth, seem to be due to the failure to fully estimate the camera's characteristics, pose, and view direction. As a result, attempts to lighten LLFF's data through Extreme View Synthesis failed.

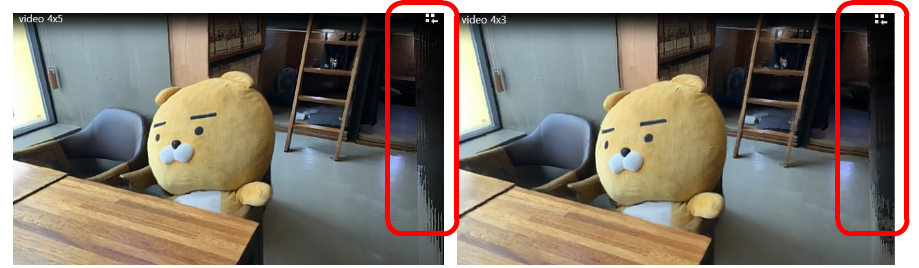
**Discussions**

We aimed to reduce the weight of data from Novel view Synthesis in general. The first attempt is to replace the frame interpolated frame through RRIN used in video interpolation. As a result, our interpolated-dataset showed particularly high results in SSIM, which evaluates human visual quality differences, compared to 4x4 or 4x3 LLFF.

For Ryan object, the quantitative evaluation value of 4x3 was the highest because the GT we assumed to match the view direction of the comparator was the result of 4x5 LLFF. The same edge-artifacts occurred due to the limitations of the LLFF model as shown below, and it is estimated that the results were high in quantitative evaluation to measure the similarity between them.

**Figure 10.**

*Specific Frame Comparison of 4x5 LLFF and 4x3 LLFF*



Moreover, we agonized over whether we could make our data lighter with extra novel-view synthesis, which is called Extreme View Synthesis. Using this model, we found that the depth-map estimation used to create a Novel view did not work well due to poor quality.

To solve this problem, we changed the depth-map estimation used in the intermediate stage to DPSnet to predict that view synthesis will be possible through a better depth-map. In the process, the depth-map itself showed very improved results as shown in Figure 8. In Figure 9, there is an artifact, but the result of view synthesis was found. However, it was too much to apply this directly to LLFF. This is because the camera parameters required for view synthesis could not be fetched into a photo with a long distance between views like our input data.

In conclusion, our study attempted to lighten LLFF by producing higher-quality data than the lightweight data with frame interpolation after extreme view synthesis, but unfortunately it did not succeed until that point. However, it is encouraging in that it attempted to improve the performance of depth maps and not only frame interpolation but also extra view interpolation. However, what is regrettable is that it would have been better if the data set was more refined.

There were many transparent features of the campus object we were trying to shoot. In this case, it is unsuitable for view synthesis. Also, the quality of the data we used was not good. Several attempts made the data back into camera parameters, but at this time, the difference between the size of the object and the view point between the data taken was not deeply considered. In the case of Extreme view synthesis, it is judged that the baseline between dataset views is too wide and does not work well. It might be produced comparable performance if I took the focal used in the original text of LLFF and took the data as it was.

**References**

Choi, I. (2019). Extreme View Synthesis, *ICCV 2019*, <https://doi.org/10.48550/arXiv.1812.04777>

Li, H. (2020). Video Frame Interpolation Via Residue Refinement, *ICASSP 2020*, <https://doi.org/10.1109/ICASSP40776.2020.9053987>

Mildenhall, B. (2019). Local Light Field Fusion: Practical View Synthesis with Prescriptive Sampling Guidelines*, SIGGRAPH 2019*, 38(4), <https://doi.org/10.48550/arXiv.1905.0088>

Mildenhall, B. (2020). NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis*, ECCV 2020*, <https://doi.org/10.48550/arXiv.2003.08934>

Murtagh, F. (1991). Multilayer perceptrons for classification and regression*, Neurocomputing*, <https://doi.org/10.1016/0925-2312(91)90023-5>

Niemeyer, M & Geiger. A. (2020). GIRAFFE: Representing Scenes as Compositional Generative Neural Feature Fields, *CVPR 2021*, <https://doi.org/10.48550/arXiv.2011.12100>

O`Shea. K & Nash. R. (2015). An Introduction to Convolutional Neural Networks*, CVPR 2015*, <https://doi.org/10.48550/arXiv.1511.08458>

Samavati, T & Soryani, M. (2023). Deep learning-based 3D reconstruction: a survey, *Artificial Intelligence Review 2023*, <https://doi.org/10.1007/s10462-023-10399-2>

Xia, W. (2021). GAN Inversion: A Survey, CVPR 2021, <https://doi.org/10.48550/arXiv.2101.05278>