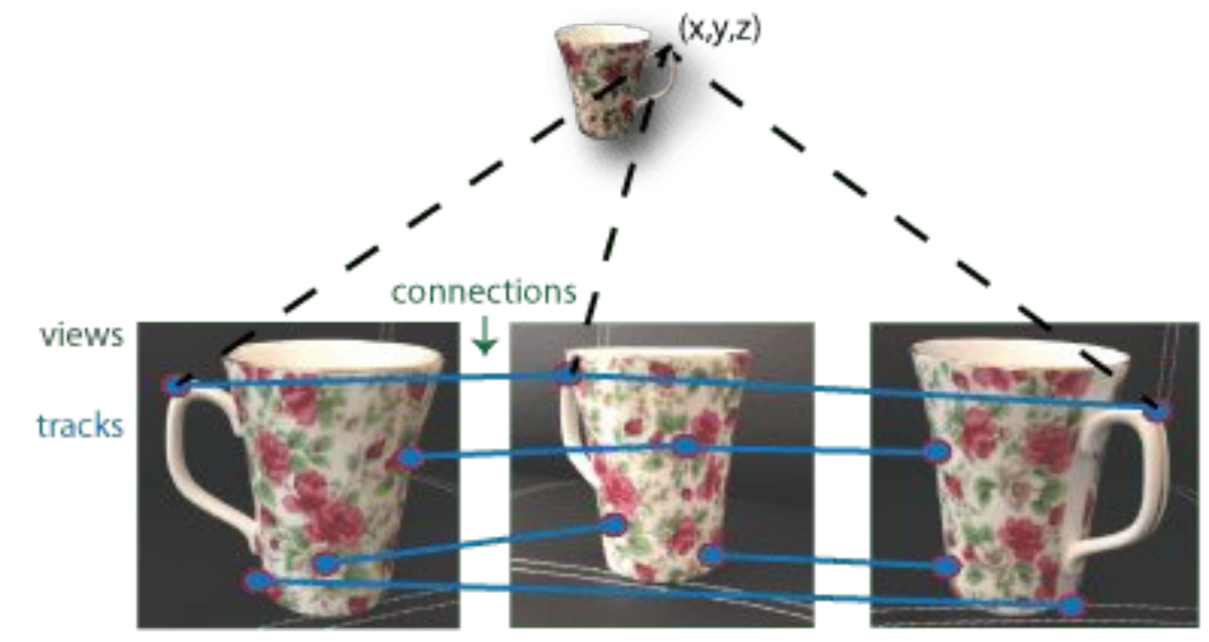
Structure from motion for 3D object model construction

Nadeen Eslam Mohamed(*Media Engineering Technology*) and **Prof. Mohammed Salem** (*Digital Media Engineering*)
Email:nadeeneslam00@gmail.com

NeRF

Representing real-world images through computers is a very challenging topic as most of the images we see on the computer are 2D images. We have to imagine the 2D pictures in our heads in order to get a possible view of how it looks like if it were 3D. In that topic, we do not have to use our imagination anymore. We could just collect 2D pictures and use a system to see how it looks like in 3D

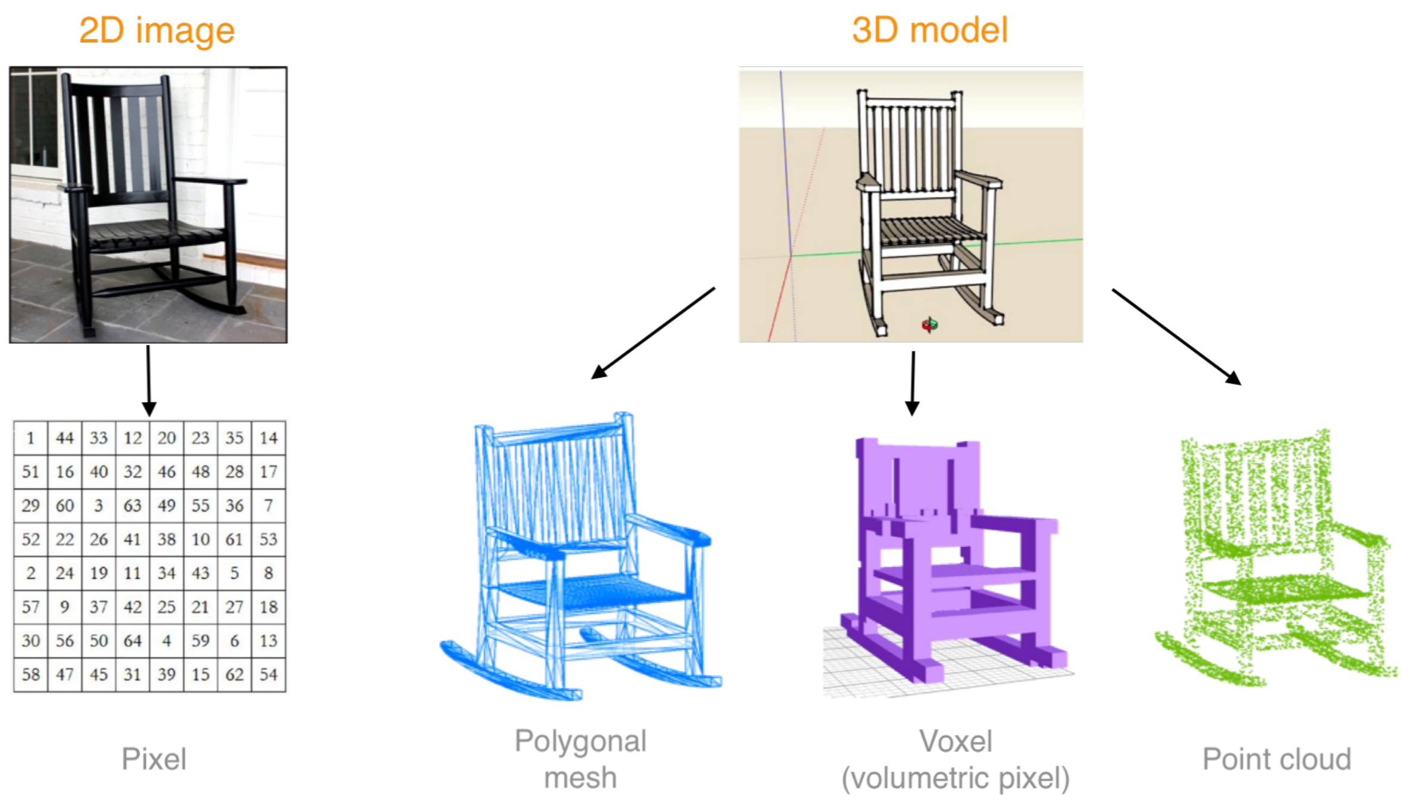


Literature Review

In short NeRF is an implicit MLP¬based model that maps 5D vectors (3D coordinates plus 2D viewing directions) to output volume density and view dependent emitted radiance at that spatial location, using fully connected (non-convolutional) deep network, computed by fitting the model to a set of training views.
One of the first studies that was carried was the Structure from motion
This approach works by matching the similar features and storing them together. Also , determining the 3D location of points within a scene and then triangulation is used to construct the 3D scene. Advantages of the SFM approach is that is matches the keypoints from different pictures and combine it, making the dataset easier when working with it. However, one of the drawbacks is that the user has much less involvement in data quality control and the origins of error in data may not be identifiable. That study brought up the topic of the NeRF as we want to obtain 3D outputs with high resolution , less rendering and training time. [2]
[3] The first study represents the NeRF and how it was firstly implemented back in 2020. For the input, they used set of 2D pictures for an object taken by a camera from different perspectives. For a single view point, march camera rays through the scene to generate a sampled set of 3D points and then use those points and their corresponding 2D viewing directions as input to the neural network to produce an output set of colors and densities, thus the input images are represented using 5D points , a 3D location x = (x, y, z) and 2D viewing direction (θ , ϕ), and whose output is an emitted color c = (r, g, b) and volume density σ (while allowing the RGB color c to be predicted as a function of both location and viewing direction)as represented in Fig. 1 . One of the advantages of that method is that it decreased the storage cost of calculating voxel grids when modeling complex scenes and also deep train networks to predict sampled volumetric representations faster.
On the other hand, the disadvantages of that implementation are that the output resolution is low as the input components are not sufficient enough in representing a high-quality output. Also, it takes at least 12 hours to train per scene.According to the first study, a lot of limitations have been pointed out, thus another study has been made to modify the NeRF in order to have better outputs.[4] This modification is called the NeRF++ , they found that the normal NeRF fails when the images are put in incorrect geometry, thus they'll try to adjust the geometry of the pictures in that study in order to avoid any ambiguity and obtain better results. In addition to that, the NeRF fails when there is a unbounded scene, thus they'll try to fix that issue in this study. Those two reasons were the main motivation behind carrying out that new study.

Problem Statement

In my research I will be developing a system improving the features of the NeRF, in which it takes 2D pictures and returns an output in less time and with better quality. In that alternative, we will try to decrease the training and rendering time of the NeRF and also obtain better quality. The image below shows how the 2D image is converted to 3D model using three perspectives: Polygonal Mesh , Voxel and point cloud



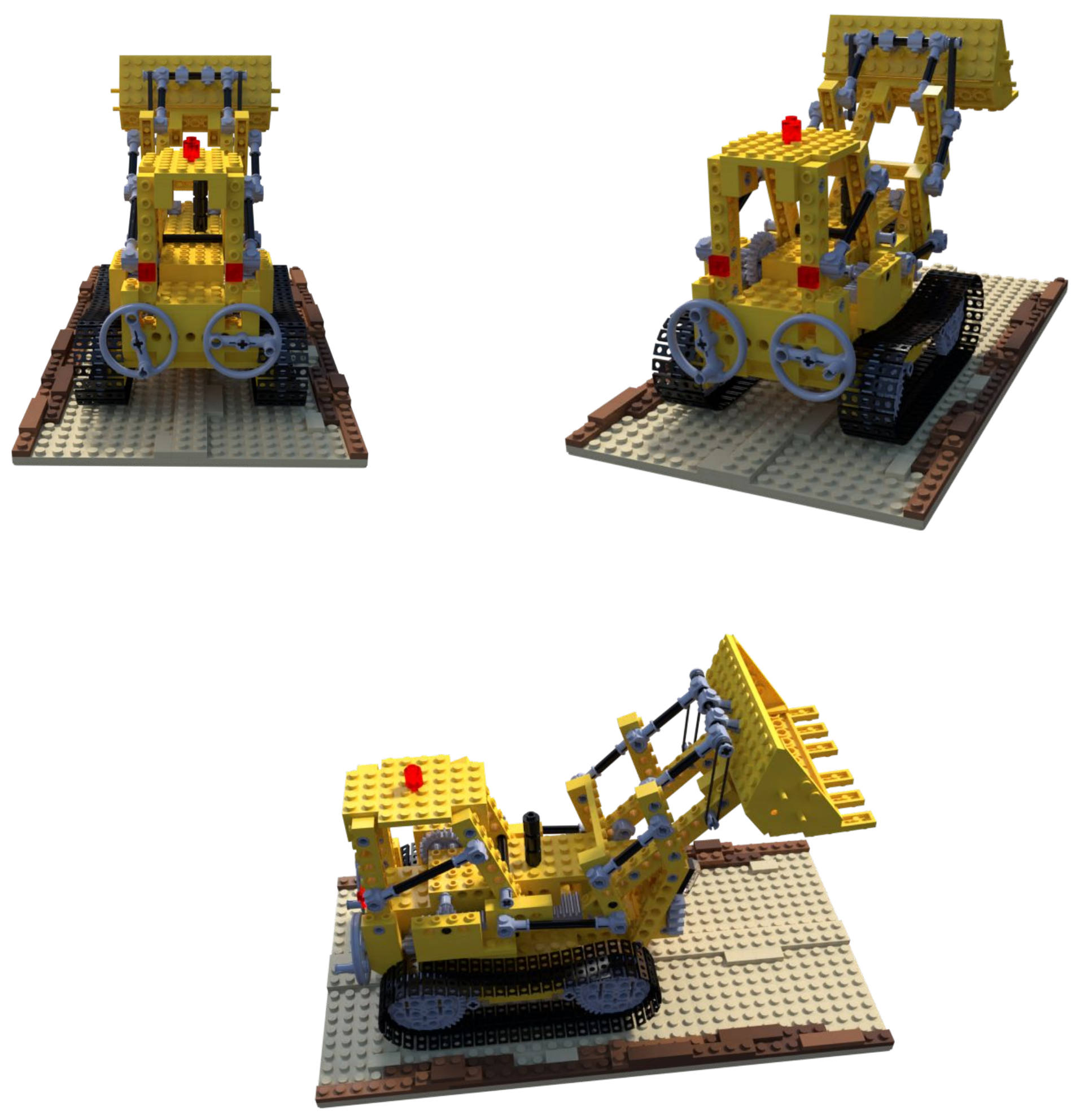
1	44	33	12	20	23	35	14
51	16	40	32	46	48	28	17
29	60	3	63	49	55	36	7
52	22	26	41	38	10	61	53
2	24	19	11	34	43	5	8
57	9	37	42	25	21	27	18
30	56	50	64	4	59	6	13
58	47	45	31	39	15	62	54

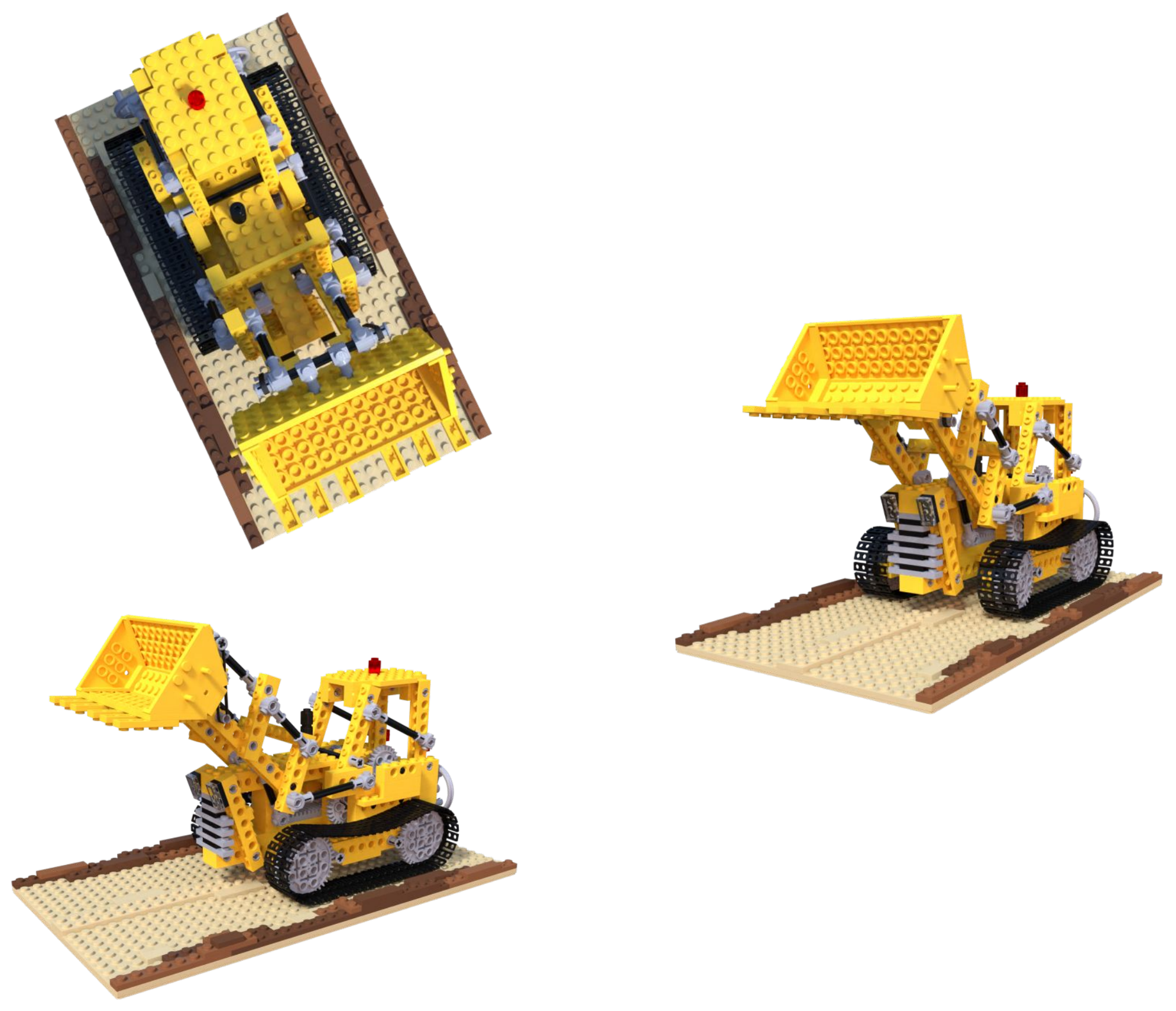
Methodology

In this part, the system will be implemented on several modules:
First module : website
1)implement a website where the user can upload several 2D pictures with a friendly user interface
2) The user will get an email when the output is ready and can download it

Second module : NeRF
1)CUDA kernels for sampling, empty space skipping, early ray termination, positional encoding, network evaluation and alpha blending
2)train both the NeRF baseline and KiloNeRF for a high number of iterations to find the limits of the representation capabilities of the respective architectures.
3)[4] uses GANs which helps in classifying the models into two sub-models , either real or fake . by implementing generator and discriminator

Datasets





Results

The result of the 3D model of the datasets shown is represented as a QR Code below in order to be able to view the 3D model.
kindly scan the QR code using the mobile camera.

References

[1] Nyimbili, P. H. Demirel, Şekerand D.Z, and Erden, "Structure from motion (sfm) –approaches applications," 9 2016.

[2] B. Mildenhall, P. P. Srinivasan, M. Tancik, and R. R. Jonathan T. Barron, "Nerf: Representing scenes as neural radiance fields for view synthesis," 8 2020.

[3] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthn, and R. Ng,
"Nerf: Representing scenes as neural radiance fields for view synthesis,"
COMMUNICATIONS OF THE ACM, vol. 65, 1 2022.

[4] K. Zhang, G. Riegler, N. Snavely, and V. Koltun, "Nerf++: Analyzing and improving neural radiance fields," 10 2020.

[5] N. Paul, "Improving neural radiance fields using transfer learning for efficient scene reconstruction," 10 2021

Prepared for Thesis Poster Display Conference

11th -12th June 2022

Faculty of Management Technology
THESIS Poster Display Conference 2022