EQUIVARIANT SCENE TRANSFORMER FOR NOVEL VIEW SYNTHESIS

Ufuk Umut Şentürk
Department of Computer Engineering
Hacettepe University
Ankara, Turkey
n19133683@cs.hacettepe.edu.tr

ABSTRACT

Novel view synthesis without 3D ground truth of the scene is ill-posed problem and requires image supervision. Therefore, having ground truth does not always give best results such that we need to enforce some important elements about 3D transformations and consistencies. In this paper, we propose novel transformer architecture enabling equivariant properties of the scene with effective usage of positional encoding and camera view encoding. There we try to explain relation between coordinate-based approaches with transformer architecture. Even though, we cannot fully outperform baseline and provide some ideas about them.

1 INTRODUCTION

Novel view synthesis(NVS) is an active area of research for both computer graphics and computer vision communities. This brings different perspectives together to solve such a problem needs to be solved for area of robotics, vision based autonomous tasks. We represents this problem as a 3D scene representation learning our transformer architecture generates and renders into the image from unseen viewpoint during training. Neural scene representation is proposed by Sitzmann et al. (2019b) and Mildenhall et al. (2020) cooperating with rendering algorithms heavily studied in computer graphics over the years. We need to consider many variations inherited by 3D scene such textures, lightning conditions etc. which are more hard to solve than the 2D image cases.

Our algorithm requires no explicit 3D ground truth and is designed to use 3D transformation bias to exploit equivariant representations as in explicit representations such as voxels, point clouds, meshes. Equivariant property of the transformations provides consistencies that constraints solution space and stabilize training. Therefore, we need relative transformation between different camera frames. Our model is based on Transformer variant as in Dosovitskiy et al. (2021), however we are the first to use full transformer architecture and equivariant positional encoding which we hope to develop at the final paper. Once we train model, we can generate novel images from novel viewpoint using only input image and relative 3D transformation in the same scene.

ShapeNet Chang et al. (2015) is used for benchmarking even its basic object centric nature providing us the sanity check of the model. Therefore, two datasets, MugsHQ and moutains, introduced by Dupont et al. (2020) will be used for complex and natural scenes as in real life. Our contributions are; full transformer model for scene representation task, equivariant positional encoding module for better convergence.

2 RELATED WORKS

Neural Rendering is another way to generate images from noisy, incomplete images, explicit or implicit scene representations. Especially, VoxNet Maturana & Scherer (2015) uses voxel grids, Thies et al. (2019) uses incomplete 3D inputs to convert scene representation into implicit neural texture representation.

Sitzmann et al. (2019b) generates images by marching rays for all pixels with LSTM model producing ray step to march ray into the geometry of the scene. Therefore, Mildenhall et al. (2020) uses

similar coordinate-based implicit neural representation of the scene encoding via basic multi-layer perceptrons which is function of ray direction and sampled 3D location on the ray. Therefore. They renders images with volumetric rendering algorithm Drebin et al. (1988). Another type of implicit representation is multi-plane images Zhou et al. (2018) which can be seen discretized version of the NERF, which warps 2D planes ordered along depth according to their corresponding depth position. Those planes represent scene properties between its current depth and previous plane which discretize 3D volume. DeepVoxels Sitzmann et al. (2019a) is grid based model having similar architectural choices as in HoloGAN Nguyen-Phuoc et al. (2019) encoding scene into latent 3D embeddings, both are CNN analogous of our transformer architecture. Most similar method to our method is Equivariant Neural Rendering Dupont et al. (2020) which introduces equivariant transformation. However, our model is based on transformer, exploiting consistencies between grid representations such that they must be same in the same scene providing that we can decompose scene into light direction dependent diffuse color part, and specular effect part.

Equivariance propoperty of the models are studied rotations by Cohen & Welling (2016) firstly by rotation filters for discrete rotations. Transformable bottleneck networks Olszewski et al. (2019) also uses similar approach as ours however does not use 3D transformations. Therefore, Equivariant Transformer Networks Tai et al. (2019) also learns equivariance properties from single image by extending Spatial Transformer Jaderberg et al. (2015). Note that transformer concept does not point out attentional transformer model. Besides, translational equivariance is basic property of the convolutional neural networks, thus we need to use positional encodings to have similar properties.

Even transformers are introduced in natural language processing, Vision Transformer Dosovitskiy et al. (2021) very recently shows that transformer models are state-of-the art for image recognition task. There are many variations of the Vision Transformer for videos Arnab et al. (2021) using 3D transformer versions corresponding 3D convolutional layers, depth estimation Ranftl et al. (2021), and also novel view synthesis tasks such as Wang et al. (2021) which uses fully transformer architecture however it is multi-view unlike ours single view architecture and also uses learnable 3D view embeddings. Other method using Transformer architecture is Rombach et al. (2021) based on VQ-GAN which is variant of VQVAE Razavi et al. (2019), however it uses SynSin Wiles et al. (2020) neural renderer as backbone and Transformer as autoregressive model on distribution.

3 METHOD

Equivariant property of rotation transformation is defined as below;

$$T^Z g(z) = g(T^Z z) \tag{1}$$

where T^Z and T^X are equivariant transformations and g is image generation function. It means rotating image with this transformation and applying function g is equivalent to generating image with function g before transforming mesh and applying transformation later.

Thus, using this property we design Siamese training similar to Dupont et al. (2020). Therefore, decomposing specular and ambient parts of the scene enforces lightning bias of the same scene such that we can model non-Lambertian effects separately .

Figure 1 shows full architecture. It has 2D-3D-2D flow for processing images similar to Nguyen-Phuoc et al. (2019). However, we replace 3D convolutional layers with 3D transformers. 3D transformers only differs while creating patches such that more patches are applied. Note that we can change depth of transformers much as resources allow. Our model leverage equivariant property of transformation such that predicting image A from input image B while transforming 3D scene features with transformation between image A and image B.

Therefore, vanilla attention is quadratic in terms of sequence length which is hard to train in practice. We changed vanilla transformer with Nyströmformer Xiong et al. (2021) approximating self-attention with linear complexity. It is based on low-rank matrix approximation called Nsyström method. Thus, ıt uses landmark technique to avoid same QK^T quadratic computation complexity. We select $\widetilde{K}$ and $\widetilde{Q}$ called landmarks from columns of K and V corresponding to Key and Value matrices as below vanilla attention;

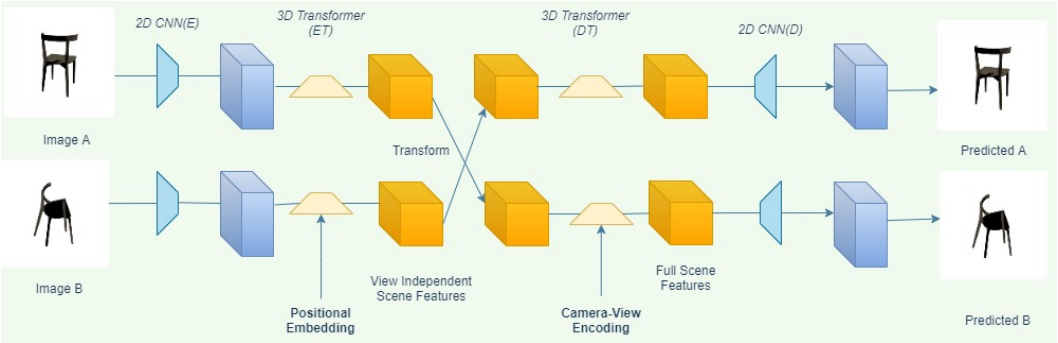

Figure 1: Proposed model, it transforms 3D scene features as rigid body to another view. Note that this model will be changed as we solve problems we face during development.

$$S = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \tag{2}$$

where d_k is number of keys. Thus we use Nyströmformer;

$$S \approx S_{Q\widetilde{K}} S^{+}_{\widetilde{Q}\widetilde{K}} S_{\widetilde{Q}K} \tag{3}$$

where landmarks are computed from averages over predefined keys and queries which segment-means approach and $+$ denotes Moore-Penrose inverse computed via Singular Value Decomposition.

We use learnable 1D positional embedding and 1D fixed camera view encoding which is similar to fixed positional encoding of elevation and azimuth angles of camera. This design choice is based on coordinate-based method such as Mildenhall et al. (2020) using similar encoding as proposed in Vaswani et al. (2017) which we believe that we can model same kind of behavior with transformers providing us model without complex computer graphics rendering algorithms. Fixed encodings are calculated as below;

$$PE\left(angle, 2i\right) = \sin\left(\frac{angle}{10000^{2i/d}}\right), PE\left(angle, 2i+1\right) = \cos\left(\frac{angle}{10000^{2i/d}}\right) \tag{4}$$

where i is frequency index in embedding dimension, angle is elevation or azimuth which are same for each element and d is dimension of embedding element.

3.1 Loss

We use L1 loss between predicted and input images between each other;

$$L_{image} = ||x_1 - f\left(x_2\right)|| + ||x_2 - f\left(x_1\right)|| \tag{5}$$

where f is out network, x_1 and x_2 are input images which we try to predict from each other.

Therefore we use SSIM loss;

$$L_{SSIM} = 1 - SSIM \tag{6}$$

where SSIM is metric known as Structural Similarity Index Wang et al. (2004). We also try to leverage consistency such that views from same scene have same view-independent lightning. This way, we try to achieve decomposition of properties better providing us better convergence. As in Figure 1, scene features are transformed to each others camera space thus transformed scene features from Image B should be same as before transformation scene features of image A. Thus,

$$L_{scene} = ||E(x_1) - T_{x_2 \to x_1}(E(x_2))|| + ||E(x_2) - T_{x_1 \to x_2}(x_1)|| \quad (7)$$

where E is encoded scene features before transformation and T is transformation function between views x_1 and x_2. Note that we did not employ this term immediately because scene features are not initially good enough to be proxy ground truth to each other, that is why we add this term after some epoch.

We also considers adversarial loss and feedback loss such that using predicted images to extract similar features of corresponding input views. Thus far, loss function is as below;

$$L = \lambda_1 L_{image} + \lambda_2 L_{SSIM} + \lambda_3 L_{scene} \quad (8)$$

where λ_1, λ_2, λ_3 are weighting hyperparameters of losses.

4 EXPERIMENTS

4.1 DATASET

We evaluate our method Mountain, MugsHQ (Dupont et al., 2020) and, ShapeNet (Chang et al., 2015) dataset. PSNR values are calculated for each dataset to quantify synthesis performance. We use images with 128x128 resolution as network inputs and produce $64x32x32x32$ 3D scene features. ShapeNet dataset is object-centric and simple dataset that we can use check for model sanity. Therefore, as in Dupont et al. (2020), we evaluate our model with photo-realistic MugsHQ and Mountain dataset consisting of satellite images of landscapes.

Table 1: Architecture Details

Architecture Type	**Head**	**Patch/Filter Size**	**Input Shape**	**Output Shape**
2D CNN	-	3	(3x128x128)	(128x32x32)
2D CNN	-	3	(128x32x32)	(1024x32x32)
3D Transformer	4	1	(32x32x32x32)	(64x32x32x32)
3D Transformer	4	1	(64x32x32x32)	(32x32x32x32)
2D CNN	2	3	(1024x32x32)	(256x32x32)
2D CNN	1	3	(256x32x32)	(3x128x128)

4.2 IMPLEMENTATION AND DETAILS

First, we implement our proposed model in Figure 1 which we call "Equivariant Transformer Renderer". We use $\lambda_1 = 0.9$, $\lambda_1 = 0.05$, $\lambda_3 = 0.1$. You can see details of architecture details in Table 1. We train our model 100 epochs with Adam optimizer using 0.02 learning rate on Mugs and Mountain dataset. Unfortunately, chair and cars dataset are not finished training such that we use models from 50. epoch and 40. epoch respectively. We have tried Nyström attention with other linear attention mechanisms and 2D+1D version of TimesFormer. However, TimesFormer version still uses quadratic time and changing its attention mechanism to linear still and other attention mechanisms do not outperform Nyström attention. Unfortunately, our ablation study is not complete.

Table 2: PSNR results of methods on test set.

Model	**Chairs**	**Cars**	**Mugs**	**Mountain**
ENR	22.83	22.26	25.98	15.48
Ours	20.79	20.52	18.75	15.54

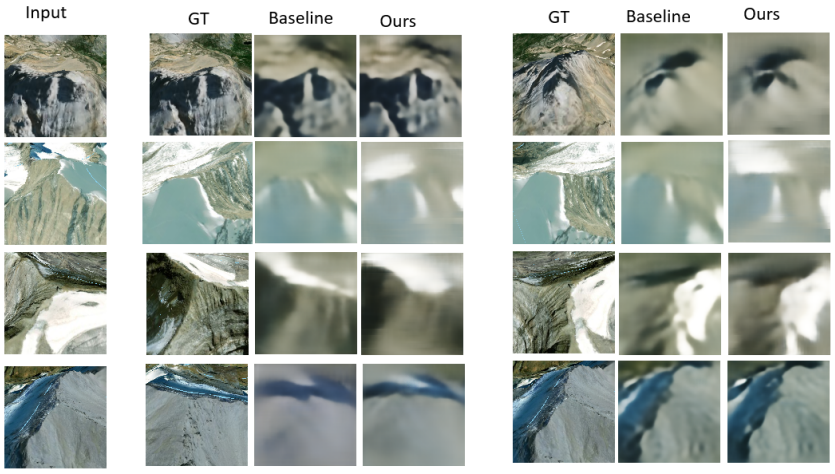

Figure 2: Outputs of ETR(ours) and ENR(baseline) models trained on Mountains dataset.

4.3 RESULTS

As you can see Figure 5, we might obtain similar or better results in the chairs and cars dataset and ENR model gives eligible outputs compared to ours in Mugs dataset. However, we correct our model giving misalignment artifacts such that image rows and columns do not align with each other in Figure 6. We infer that it is because the vanilla positional embedding causing such as having hard time encoding 3D spatial positions from positional embeddings propagated from 2D transformers. Then, we apply positional encoding to the 3D Transformer and replace 2D transformer with 2D CNNs to have convolutional bias and also better run time.

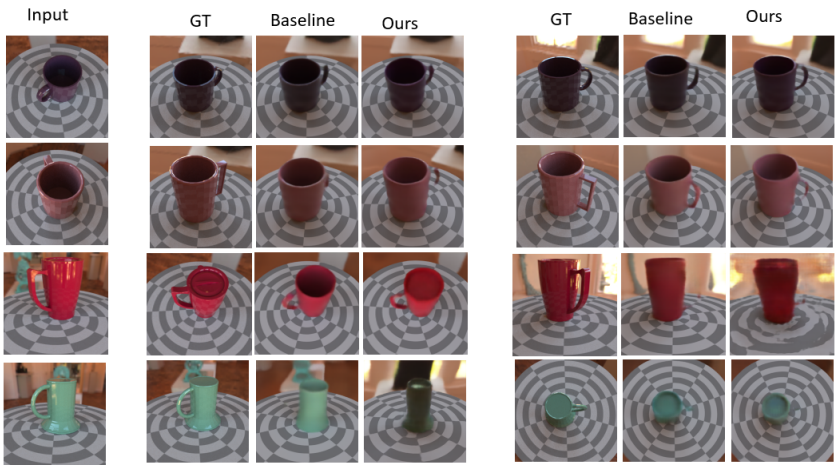

Figure 3: Outputs of ETR(ours) and ENR(baseline) models trained on MugsHQ dataset.

As you can see in Figure 3, there are different cases from baseline. At the first row, our model generates better images than baseline in terms of details and structure. After the second row, you can see table under the mugs does not rotate correctly as in the baseline case and model gets confused about second example at third row. This bothers me that might be the case learnable positional encodings could not generalize for those samples. As stated before, we could not ablate model very well, however we know that previous artefact caused by this particular usage. Thus we should have run model with fixed positional encoding and model with transformed positional encoding for the second 3D transformer. Furthermore as you can see the third row, we can generate closed cup even

the hint is very little coming from input even baseline generates image with open cup. This main problem particularly is the reason of low performance quantitatively.

You can see scene captured from far distance are not generated very well in Figure 2. Generally so much angle difference generates blurrier results. First sample at the first row and second sample at the last row are same as input and we can see that those results are better than images generated by other angles. As you can see PSNR results on Table 2, we just only outperform baseline in this dataset with small margin.

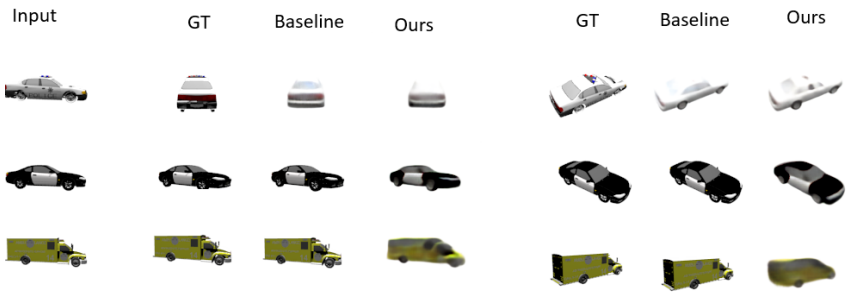

Figure 4: Outputs of ETR(ours) and ENR(baseline) models trained on ShapeNet Cars dataset. Note that our model is not trained 100 epochs as baseline, only 50 epochs.

For the ShapeNet Car dataset, our model in not completely trained and produces worse results than baseline as expected. There are still bias on the cars in Figure 4 such that given truck image as input produces car like shape on different viewpoint meaning poor generalization.

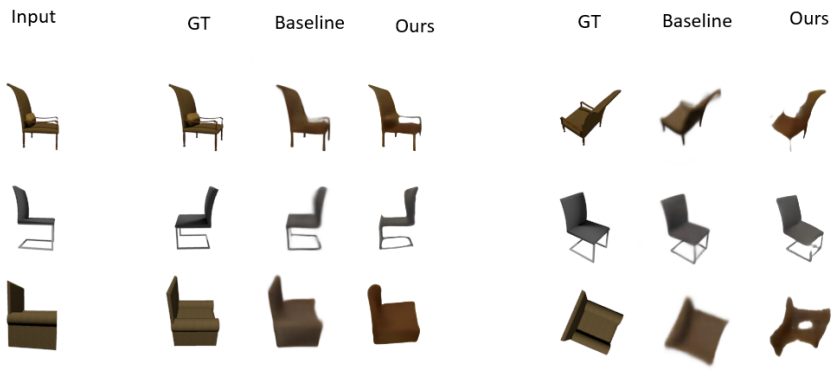

Figure 5: Outputs of ETR(ours) and ENR(baseline) models trained on chairs dataset.Note that our model is not trained 100 epochs as baseline, only 40 epochs.

For the ShapeNet Chairs dataset, we also cannot train model completely, only 40 epochs. However, in Figure 5, we see that our model generates less blurry images compared to baseline. Since it is not converged fully, we see that there is viewpoint bias such that it generates good images when viewpoint difference is not much as we expected. This can be overcome with more training time.

5 Conclusion

Our work has its limits and wide range of future work. As mentioned before, we might use different positional encoding strategy and we need to do more extensive ablation study. However, we can say that our model extract equivariant transformation scene features and even using little hints to sythesis images based on novel viewpoint. Furthermore, It still needs tuning such that we can outperform this baseline with same run-time with baseline which is one of the keypoint of the method.

Figure 6: Misalignment artifacts. Left side is ground truth, right one is our old output where we use all transformer architecture.

REFERENCES

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer, 2021.

Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015.

Taco Cohen and Max Welling. Group equivariant convolutional networks. In Maria Florina Balcan and Kilian Q. Weinberger (eds.), *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pp. 2990–2999, New York, New York, USA, 20–22 Jun 2016. PMLR. URL http://proceedings.mlr.press/v48/cohenc16.html.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

Robert A. Drebin, Loren Carpenter, and Pat Hanrahan. Volume rendering. In *Proceedings of the 15th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '88, pp. 65–74, New York, NY, USA, 1988. Association for Computing Machinery. ISBN 0897912756. doi: 10.1145/54852.378484. URL https://doi.org/10.1145/54852.378484.

Emilien Dupont, Miguel Bautista Martin, Alex Colburn, Aditya Sankar, Josh Susskind, and Qi Shan. Equivariant neural rendering. In Hal Daumé III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 2761–2770. PMLR, 13–18 Jul 2020. URL http://proceedings.mlr.press/v119/dupont20a.html.

Max Jaderberg, Karen Simonyan, Andrew Zisserman, and koray kavukcuoglu. Spatial transformer networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper/2015/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf.

Daniel Maturana and Sebastian Scherer. Voxnet: A 3d convolutional neural network for real-time object recognition. In *Proceedings of (IROS) IEEE/RSJ International Conference on Intelligent Robots and Systems*, pp. 922 – 928, September 2015.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Thu Nguyen-Phuoc, Chuan Li, Lucas Theis, Christian Richardt, and Yong-Liang Yang. Hologan: Unsupervised learning of 3d representations from natural images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Kyle Olszewski, Sergey Tulyakov, Oliver Woodford, Hao Li, and Linjie Luo. Transformable bottleneck networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. *ArXiv preprint*, 2021.

Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/5f8e2fa1718d1bbcadf1cd9c7a54fb8c-Paper.pdf.

Robin Rombach, Patrick Esser, and Björn Ommer. Geometry-free view synthesis: Transformers and no 3d priors, 2021.

Vincent Sitzmann, Justus Thies, Felix Heide, Matthias Nießner, Gordon Wetzstein, and Michael Zollhöfer. Deepvoxels: Learning persistent 3d feature embeddings. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2019a.

Vincent Sitzmann, Michael Zollhöfer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. In *Advances in Neural Information Processing Systems*, 2019b.

Kai Sheng Tai, Peter Bailis, and Gregory Valiant. Equivariant transformer networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6086–6095. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/tai19a.html.

Justus Thies, Michael Zollhöfer, and Matthias Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Trans. Graph.*, 38(4), July 2019. ISSN 0730-0301. doi: 10.1145/3306346.3323035. URL https://doi.org/10.1145/3306346.3323035.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Dan Wang, Xinrui Cui, Xun Chen, Zhengxia Zou, Tianyang Shi, Septimiu Salcudean, Z. Jane Wang, and Rabab Ward. Multi-view 3d reconstruction with transformer, 2021.

Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *Trans. Img. Proc.*, 13(4):600–612, April 2004. ISSN 1057-7149. doi: 10.1109/TIP.2003.819861. URL https://doi.org/10.1109/TIP.2003.819861.

Olivia Wiles, Georgia Gkioxari, Richard Szeliski, and Justin Johnson. Synsin: End-to-end view synthesis from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Yunyang Xiong, Zhanpeng Zeng, Rudrasis Chakraborty, Mingxing Tan, Glenn Fung, Yin Li, and Vikas Singh. Nyströmformer: A nyström-based algorithm for approximating self-attention, 2021.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. In *SIGGRAPH*, 2018.