DOUBLE NERF: REPRESENTING DYNAMIC SCENES AS NEURAL RADIANCE FIELDS

V. V. Kniaz[1,2,*], V. A. Knyaz[1,2] A. Bordodymov[1] P. Moshkantsev[1], D. Novikov[1], S. Barylnik[1]

[1] State Res. Institute of Aviation Systems (GosNIIAS), 125319, 7, Victorenko str., Moscow, Russia –
(vl.kniaz, bordodymov, petr_mosh, daninov, semyon_barylnik)@gosniias.ru
[2] Moscow Institute of Physics and Technology (MIPT), Russia –
kniaz.va@mipt.ru

Commission II, WG II/8

KEY WORDS: nerual radiance fields, novel view synthesis, 3D scene reconstruction

ABSTRACT:

Neural Radiance Fields (NeRFs) are non-convolutional neural models that learn 3D scene structure and color to produce novel images of a given scene from a new view point. NeRFs are closely related to such photogrammetric problems as camera pose estimation and bundle adjustment. NeRF takes a number of oriented cameras and photos as an input and learns a function that maps a 5D pose vector to an RGB color and volume destiny at point. The estimated function can be used to draw an image using a volume rendering pipeline. Still NeRF have a major limitation: they can not be used for dynamic scene synthesis. We propose a modified NeRF framework that can represent a dynamic scene as a superposition of two or more neural radiance fields. We consider a simple dynamic scene consisting of a static background scene and moving object with a static shape. We implemented our DoubleNeRF model using TensorFlow library. The results of evaluation are encouraging and demonstrate that our DoubleNeRF model achieves and surpasses the state of the art in the dynamic scene synthesis. Our framework includes two neural radiance fields for a background scene and dynamic objects. The evaluation of the model demonstrates that it can be effectively used for synthesis of photorealistic dynamic image sequence and videos.

1. INTRODUCTION

Neural Radiance Fields (NeRFs) (Mildenhall et al., 2020) are non-convolutional neural models that learn 3D scene structure and color to produce novel images of a given scene from a new view point. NeRFs are closely related to such photogrammetric problems as camera pose estimation and bundle adjustment. NeRF takes a number of oriented cameras and photos as an input and learns a function that maps a 5D pose vector to an RGB color and volume destiny at point. The estimated function can be used to draw an image using a volume rendering pipeline. NeRF models are widely used for photorealistic image synthesis from novel viewpoint. Still NeRF have a major limitation: they can not be used for dynamic scene synthesis.

In this paper we propose a modified NeRF framework that can represent a dynamic scene as a superposition of two or more neural radiance fields.

We used scenes from the *SematicVoxels* dataset (Kniaz et al., 2020) to train and evaluate our DoubleNeRF model. We implemented our DoubleNeRF model using TensorFlow library. We used city scenes as the background and cars as foreground dynamic objects. We evaluate our DoubleNeRF model and baselines in terms of PSNR, SSIM, LPIPs and FID metrics. We compare synthetic images generated for novel views with real images from the dataset. The results of evaluation are encouraging and demonstrate that our DoubleNeRF model achieves and surpasses the state of the art in the dynamic scene synthesis.

We proposed a novel DoubleNeRF framework for photorealistic image synthesis from novel views. Our framework includes two neural radiance fields for a background scene and dynamic objects. The evaluation of the model demonstrates that it can be effectively used for synthesis of photorealistic dynamic image sequence and videos.

2. RELATED WORK

The problem of effective and realistic representing 3D scene is one of the key problems in computer graphics. As usual a researcher has to find reasonable balance between the speed of rendering and the quality of the rendering.

The traditional and accurate methods for creating photorealistic 3D model of a real scene are photogrammetry-based ones. Currently Structure-from-Motion (Shapiro and Stockman, 2001, Knyaz and Zheltov, 2017) and Multi View Stereo are widely used tools for 3D scene reconstruction basing on a set of images, allowing to reconstruct as large 3D scenes (Liu et al., 2023), so complex 3D objects (Knyaz et al., 2020).

The progress in means and methods of machine learning gave an impulse for applying such techniques for effective synthesizing new views of complex 3D scenes with given 3D model and a set of scene images. Fully connected neural networks for synthesis new view having a number of partial images (Mildenhall et al., 2020) were recently proposed. They are termed as neural radiance fields (NeRFs) and demonstrated high performance in task of high-resolution photorealistic rendering.

NeRF uses continuous scene representation as spatial coordinate vector (x, y, z) and viewing direction (θ, ϕ)). Basing on this 5D representation, NeRFs synthesize a new view of the scene

[*] Corresponding author

Publisher's note: Copernicus Publications has not received any payments from Russian or Belarusian institutions for this paper.

This contribution has been peer-reviewed.

represented by a set of images by directly searching the parameters, that minimize the rendering error for the given set. It was shown (Mildenhall et al., 2020) that such approach allows outperforming previous works on new views synthesizing by neural rendering.

The advantages of NeRF models have attracted a lot attention in the computer vision area in the following years and initiated the researches in various areas, such as speeding the training, improving the quality of rendering for sparse views, pose estimating by NeRF.

To improve NeRF performance in the case of sparse input views, the RegNeRF model (Niemeyer et al., 2022) uses additional depth and color regularization. It allowed RegNeRF to outperform such NeRF models as PixelNeRF (Yu et al., 2021) and Stereo Radiance Fields (SRF) model (Chibane et al., 2021), that employed features from pre-trained networks or a prior conditioning for rendering. The performance comparison was performed using DTU (Jensen et al., 2014) and LLFF (Mildenhall et al., 2019) datasets.

Instant Neural Graphics Primitives (Müller et al., 2022) model today demonstrates the state-of-the-art for NeRF models in training and inference speed. The proposed approach exploits hash encoding trained simultaneously multilayer perceptrons (MLPs) of the NeRF. Along with advanced ray marching techniques including exponential stepping, empty space skipping, sample compaction, it allowed to dramatically reduce training time comparing with baselines such as mip-NeRF (Barron et al., 2021) or Neural Sparse Voxel Fields (NSVF) (Liu et al., 2021) models.

Taking the advantages of multi view stereo in generating high quality 3D scenes and their views, from the one side, and possibility of deep multi view stereo methods to reconstruct the geometry of a scene in a short time, from the other side, Point-NeRF (Xu et al., 2022) model can generate a radiance field using neural 3D point clouds fast and with high quality. The high rendering performance of the Point-NeRF is based on aggregating neural point features near scene surfaces, in a ray marching-based pipeline.

The comprehensive analysis of NeRFs considering these models from wide variety points of view as theoretical fundamentals, existing approaches, methods, and datasets, metrics used and state-of-the-art performance can be found in the dedicated reviews (Tewari et al., 2022, Gao et al., 2022). But the most relevant to our study are researches, that not only synthesize the new view of a static, but address to scene composition with NeRF. D-NeRF model (Pumarola et al., 2021) is aimed at extending NeRF to a dynamic scenes. It allows to synthesize and render new images of rigid and non-rigid motion objects.

The model uses time as an additional input, and train the model in two main steps. Firstly, the scene is encoded into a canonical space, and, secondly, this canonical representation is maped into the deformed scene at a particular time. At both stages fully-connected networks are used.

The NeRF++ model (Zhang et al., 2020), was adapted to generate novel views for unbound scenes, by separating the scene using a sphere. The inside of the sphere contained all foreground object and all fictitious cameras views, whereas the background was outside the sphere.

3. METHOD

Our model works by combining two radiance fields n_b (Figure 1) and n_o (Figure 2) representing the background scene and the object consequently. The resulting neural radiance field configuration is presented in Figure 3.

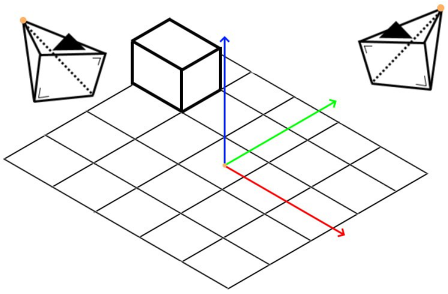

Figure 1. Camera configuration for the background neural radiance field n_b.

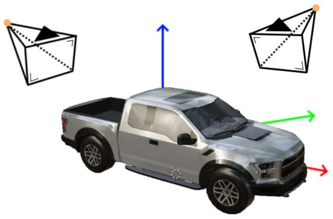

Figure 2. Camera configuration for object neural radiance field n_o.

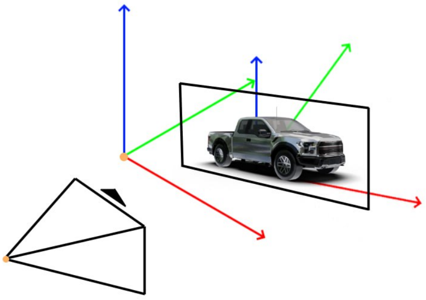

Figure 3. Superposition of two neural radiance fields.

Our contribution to the original `NeRF` model is twofold.

Firstly, we modify the object radiance field model to predict an additional transparency component α, that represents the transparency of scene at point x, y, z. We prepare the training data for the object using the alpha channel to mask the background

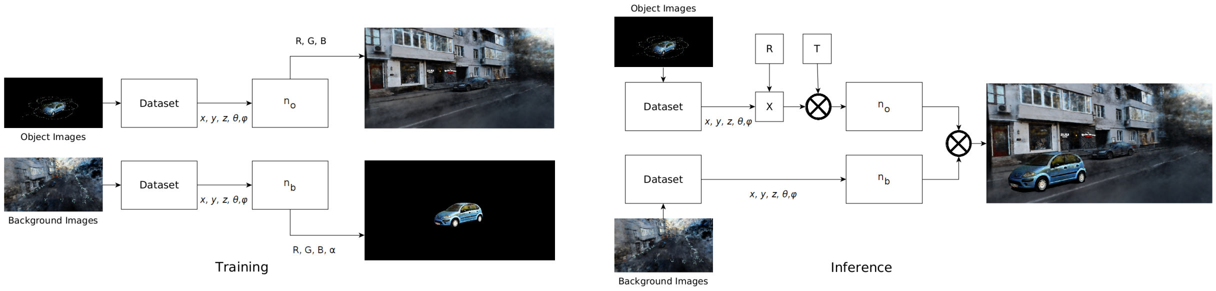

Figure 4. The `DoubleNeRF` framework overview at training (left) and inference (right) phahse

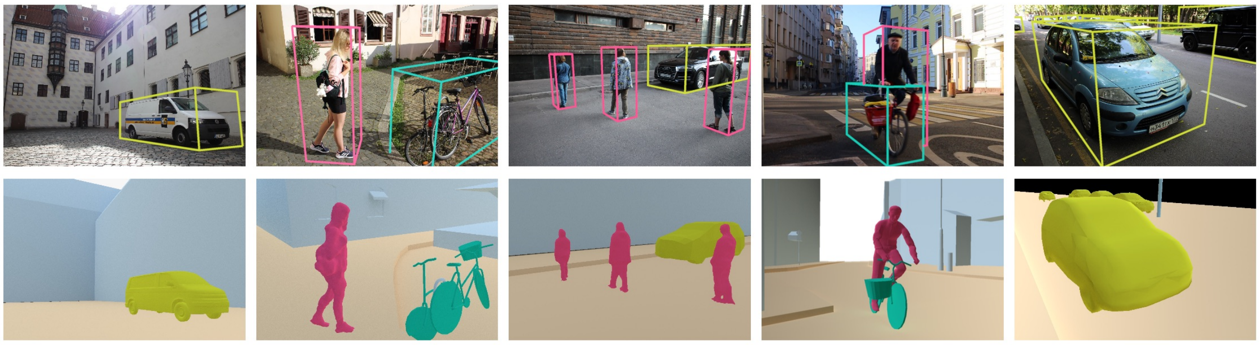

Figure 5. The examples of image scenes and corresponding voxel models from the *SematicVoxels* dataset

in the object scene. This allows us to perform 3D alpha compositing, similar to 2D splices (Kniaz et al., 2019, Kniaz et al., 2023) that are widely used for image manipulation.

Secondly, we propose to represent the resulting scene as sum of two neural radiance fields representing the static background and dynamic object.

We used scenes from the *SematicVoxels* dataset (Kniaz et al., 2020) (Figure 5) to train and evaluate our `DoubleNeRF` model.

3.1 Framework Overview

We consider a simple dynamic scene consisting of a static background scene and moving object with a static shape. We assume that the surface brightness and reflections of the moving object are independent from the location of the object with respect to the background scene. Also we assume that two sets of oriented images A_b and A_o are available.

The `DoubleNeRF` framework overview at training (left) and inference (right) phahses is shown in Figure 4.

The first set A_b is used to estimate the neural radiance field of the background scene. The set A_b does not include images of the dynamic object. The second set A_o is used to estimate neural radiance field of the dynamic object. Using these two sets, we estimate two neural radiance fields n_b and n_o. Neural radiance field n_b operates with the scene coordinate system $O_bX_bY_bZ_b$. The origin of the scene coordinate system is located in the center of mass of the scene 3D model on the ground level. The X_b axis is directed coolinear to the projection of the optical axis of the first camera in the set A_b. The Z_b is normal to the surface of the ground. The Y_b compliments the coordinate system to right-handed.

Neural radiance field n_o operates in the object coordinate system $O_oX_oY_oZ_o$. The origin of the object coordinate system is located on the projection of the center mass to the ground plane. The X_o axis is directed toward to the positive direction of the construction axis of the object (e.g., toward the forward motion of the car). The Z_o axis is normal to the surface of the ground. The Y_o axis compliments the coordinate system to right-handed.

The object neural radiance field n_o does not include background scene. In other words, for any point in n_o that is not located on the object, the volume density is equal to 0. Therefore, we can assume that the resulting dynamic radiance field is the sum of two static radiance fields:

$$n_d(x, y, z, \Theta, \phi) = n_b(x, y, z, \Theta, \phi) \\ + n_o(x'_o, y'_o, z'_o, \Theta, \phi),$$

where x'_0, y'_0, z'_0 are object coordinates transformed from the scene coordinate system to the object coordinate system,

$$X_o = [x_o, y_o, z_o]^T, \tag{1}$$

$$X'_o = R_{bo} \cdot X_o + T_{bo}, \tag{2}$$

where R_{bo} is the rotation matrix that defines a transformation from the background scene coordinate system to the object coordinate system, T_{bo} is the translation from $O_bX_bY_bZ_b$ to $O_oX_oY_oZ_o$.

3.2 Dataset Generation

We used scenes from the *SematicVoxels* dataset (Kniaz et al., 2020) to train and evaluate our `DoubleNeRF` model.

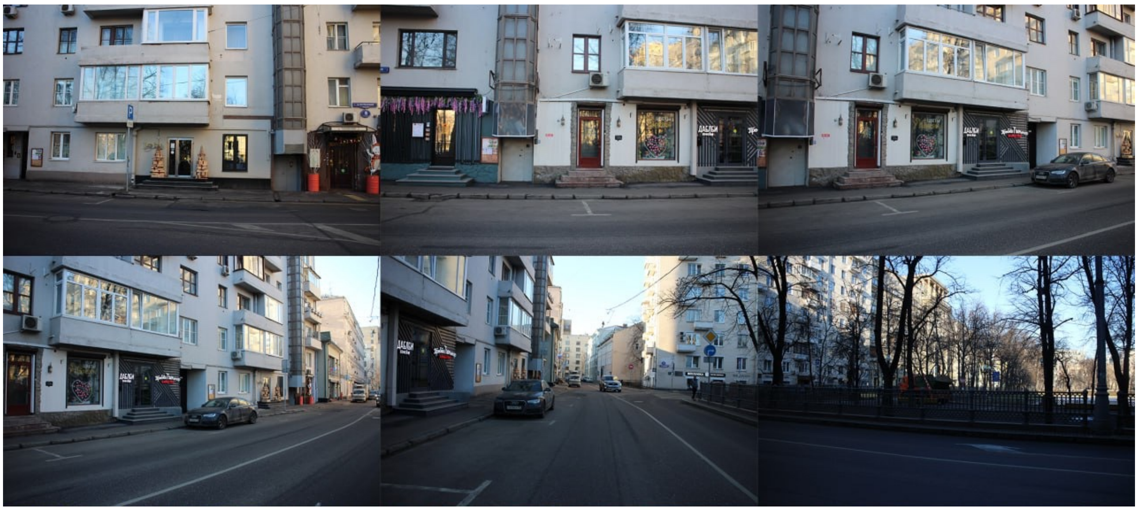

Figure 6. Example images from the training set for background scene (top) and object scene (bottom).

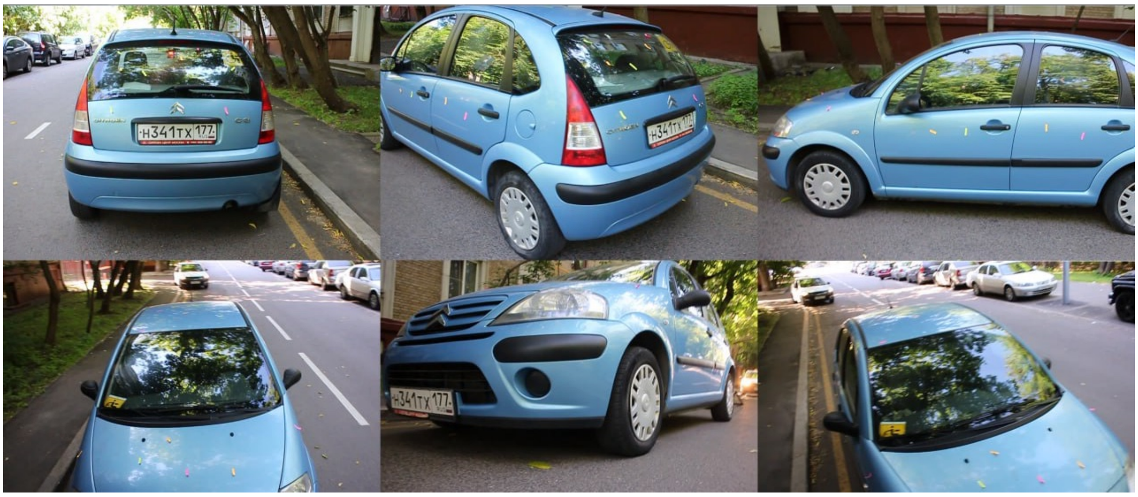

Figure 7. Example images from the training set for background scene (top) and object scene (bottom).

The Semantic Voxels Dataset consists of 116,000 samples, that presents 3D and 2D data for 36 scenes. Each data sample represents a single camera pose, and includes a color image and a camera pose for this image, a depth map and a semantic voxel model, and an object pose annotations for all classes. The dataset is consistent with *NuScenes* dataset format (Hodaň et al., 2017). Semantic Voxels dataset has two parts: real and synthetic. The real split was generated using a Structure-from-Motion (SfM) technique similar to (Hodaň et al., 2017). It consists of 16,000 images.

The examples of images of real scenes and corresponding voxel models from the *SematicVoxels* dataset are presented in Figure 5.

Example images from the training set for background scene and object scene are presented in Figure 6 and 7. We manually labelled the background in the object split of the dataset.

4. EVALUATION

4.1 Qualitative Evaluation

We evaluate our model quantitatively using example novel views generated by our algorithm. Comparison of the scene generated using only the original `NeRF` model and our `DoubleNeRF` model are presented in Figures 8 and 9.

4.2 Quantitative Evaluation

4.2.1 Metrics. We evaluate our `DoubleNeRF` model and baselines in terms of PSNR, SSIM, LPIPs and FID metrics.

The Structural Similarity Index Measure (SSIM) is calculated on various windows of an image. The measure between two images x and y of the same size $N \times N$ is:

$$\mathrm{SSIM}(x,y) = \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)}$$

μ_x, μ_y – the pixel sample mean of x and y correspondingly;
$\sigma_x^2, \sigma_x^2, \sigma_{xy}$ – the variance of of x and y, covariance of x and y correspondingly;
c_1, c_2 – two variables intended for stabilizing the division in case of weak denominator (Nilsson and Akenine-Möller, 2020);

The Peak Signal-to-Noise Ratio (PSNR) (in dB) is defined as:

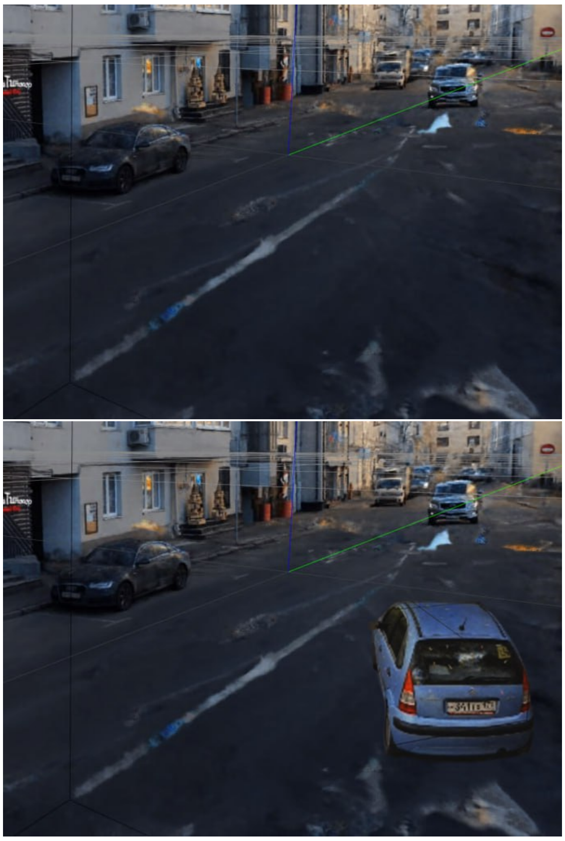

Figure 8. A novel view generated by the original `NeRF` model (top) and our `DoubleNeRF` model (bottom).

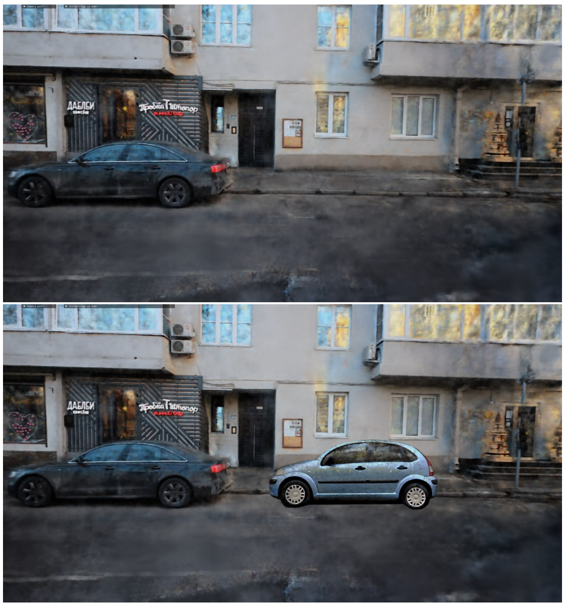

Figure 9. A novel view generated by the original `NeRF` model (top) and our `DoubleNeRF` model (bottom).

$$
\begin{aligned}
PSNR &= 10 \cdot \log_{10}\left(\frac{MAX_I^2}{MSE}\right) \\
&= 20 \cdot \log_{10}\left(\frac{MAX_I}{\sqrt{MSE}}\right) \\
&= 20 \cdot \log_{10}(MAX_I) - 10 \cdot \log_{10}(MSE),
\end{aligned}
$$

where MSE is mean squared error defined as:

$$MSE = \frac{1}{m\,n}\sum_{i=0}^{m-1}\sum_{j=0}^{n-1}[I(i,j) - K(i,j)]^2.$$

and MAX_I is the maximum possible pixel value of the image.

The Fréchet inception distance (FID) is a metric used to assess the quality of images. FID compares the distribution of generated images with the distribution of a set of real images ("ground truth"). For two multidimensional Gaussian distributions $\mathcal{N}(\mu, \Sigma)$ and $\mathcal{N}(\mu', \Sigma')$, it is given by:

$$
\begin{gathered}
d_F(\mathcal{N}(\mu,\Sigma), \mathcal{N}(\mu',\Sigma'))^2 = \|\mu - \mu'\|_2^2 + trace; \\
trace = \operatorname{tr}\left(\Sigma + \Sigma' - 2\left(\Sigma^{\frac{1}{2}} \cdot \Sigma' \cdot \Sigma^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)
\end{gathered}
$$

Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) metric is used to measure the 'perceptual' similarity between different images. It is calculated using activations of feature maps of a deep neural network (e.g., VGG (Simonyan and Zisserman, 2014)). To measure the distance between two images, each image is transformed to a feature map and L^2 distance is calculated between the correspondent feature maps. We calculate distance from synthetic images generated by a given model and real images. LPIPS measures the distance $||\cdot||$ in a CNN feature space, considering a 'perceptual loss' in an image regression problems. LPIPs can be expressed as:

$$LPIPS(x, x') := \|f_\theta(x) - f_\theta(x')\|$$

We compare our `DoubleNeRF` model with three baselines: the original NeRF model, image splice generated using the Blender 3D creation suite, and simple 2D image splice.

The results of quantitative evaluation are presented in Table 1. While the original `NeRF` model outperforms our model in all metrics it does not include an additional spliced object. Comparison with traditional 2D splicing technique demonstrates that our `DoubleNeRF` model outperforms other methods by a large margin.

	PSNR	**SSIM**	**FID**	**LPIPS**
Blender	15.0	0.712	87	0.9
2D splice	12.1	0.763	91	0.8
NeRF	40.1	0.902	23	0.25
DoubleNeRF	39.2	0.989	45	0.31

Table 1. Quantitative results for novel view synthesis.

5. CONCLUSION

We propose the `DoubleNeRF` framework for synthesizing a new view of an initial static scene described by a set of images with embedded new object, also generated by neural radiance field. A new view is generated as a superposition of two neural radiance field, and has high perceptual quality.

We compare synthetic images generated for novel views with real images from the dataset. The results of evaluation demonstrate that our `DoubleNeRF` model achieves and surpasses the state of the art in the dynamic scene synthesis.

REFERENCES

Barron, J. T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P. P., 2021. Mip-NeRF: A Multiscale Representation for Anti-Aliasing Neural Radiance Fields. *arXiv preprint arXiv:2103.13415*.

Chibane, J., Bansal, A., Lazova, V., Pons-Moll, G., 2021. Stereo radiance fields (srf): Learning view synthesis for sparse views of novel scenes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7911–7920.

Gao, K., Gao, Y., He, H., Lu, D., Xu, L., Li, J., 2022. NeRF: Neural Radiance Field in 3D Vision, A Comprehensive Review. *arXiv preprint arXiv:2210.00379*.

Hodaň, T., Haluza, P., Obdržálek, Š., Matas, J., Lourakis, M., Zabulis, X., 2017. T-LESS: An RGB-D Dataset for 6D Pose Estimation of Texture-less Objects. *IEEE Winter Conference on Applications of Computer Vision (WACV)*.

Jensen, R., Dahl, A., Vogiatzis, G., Tola, E., Aanaes, H., 2014. Large scale multi-view stereopsis evaluation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kniaz, V., Knyaz, V., Moshkantsev, P., 2023. Iq-gan: Instance-quantized image synthesis. B. Kryzhanovsky, W. Dunin-Barkowski, V. Redko, Y. Tiumentsev (eds), *Advances in Neural Computation, Machine Learning, and Cognitive Research VI*, Springer International Publishing, Cham, 277–291.

Kniaz, V. V., Knyaz, V. A., Remondino, F., Bordodymov, A., Moshkantsev, P., 2020. Image-to-voxel model translation for 3d scene reconstruction and segmentation. A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (eds), *Computer Vision – ECCV 2020*, Springer International Publishing, Cham, 105–124.

Kniaz, V. V., Knyaz, V., Remondino, F., 2019. The point where reality meets fantasy: Mixed adversarial generators for image splice detection. H. Wallach, H. Larochelle, A. Beygelzimer, F. dAlche Buc, E. Fox, R. Garnett (eds), *Advances in Neural Information Processing Systems*, 32, Curran Associates, Inc.

Knyaz, V. A., Kniaz, V. V., Remondino, F., Zheltov, S. Y., Gruen, A., 2020. 3D Reconstruction of a Complex Grid Structure Combining UAS Images and Deep Learning. *Remote Sensing*, 12(19), 3128. http://dx.doi.org/10.3390/rs12193128.

Knyaz, V., Zheltov, S., 2017. Accuracy evaluation of structure from motion surface 3D reconstruction. F. Remondino, M. R. Shortis (eds), *Videometrics, Range Imaging, and Applications XIV*, 10332, International Society for Optics and Photonics, SPIE, 200 – 209.

Liu, L., Gu, J., Lin, K. Z., Chua, T.-S., Theobalt, C., 2021. Neural Sparse Voxel Fields. *arXiv preprint arXiv:2007.11571*.

Liu, Z., Qv, W., Cai, H., Guan, H., Zhang, S., 2023. An Efficient and Robust Hybrid SfM Method for Large-Scale Scenes. *Remote Sensing*, 15(3). https://www.mdpi.com/2072-4292/15/3/769.

Mildenhall, B., Srinivasan, P. P., Ortiz-Cayon, R., Kalantari, N. K., Ramamoorthi, R., Ng, R., Kar, A., 2019. Local Light Field Fusion: Practical View Synthesis with Prescriptive Sampling Guidelines. *arXiv preprint arXiv:1905.00889*.

Mildenhall, B., Srinivasan, P. P., Tancik, M., Barron, J. T., Ramamoorthi, R., Ng, R., 2020. Nerf: Representing scenes as neural radiance fields for view synthesis. A. Vedaldi, H. Bischof, T. Brox, J.-M. Frahm (eds), *Computer Vision – ECCV 2020*, Springer International Publishing, Cham, 405–421.

Müller, T., Evans, A., Schied, C., Keller, A., 2022. Instant Neural Graphics Primitives with a Multiresolution Hash Encoding. *ACM Trans. Graph.*, 41(4). https://doi.org/10.1145/3528223.3530127.

Niemeyer, M., Barron, J. T., Mildenhall, B., Sajjadi, M. S. M., Geiger, A., Radwan, N., 2022. Regnerf: Regularizing neural radiance fields for view synthesis from sparse inputs. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5480–5490.

Nilsson, J., Akenine-Möller, T., 2020. Understanding SSIM. *arXiv preprint arXiv:2006.13846*.

Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F., 2021. D-nerf: Neural radiance fields for dynamic scenes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10318–10327.

Shapiro, L., Stockman, G., 2001. *Computer Vision*. Prentice Hall.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tewari, A., Thies, J., Mildenhall, B., Srinivasan, P., Tretschk, E., Wang, Y., Lassner, C., Sitzmann, V., Martin-Brualla, R., Lombardi, S., Simon, T., Theobalt, C., Niessner, M., Barron, J. T., Wetzstein, G., Zollhoefer, M., Golyanik, V., 2022. Advances in Neural Rendering. *arXiv preprint arXiv:2111.05849*.

Xu, Q., Xu, Z., Philip, J., Bi, S., Shu, Z., Sunkavalli, K., Neumann, U., 2022. Point-nerf: Point-based neural radiance fields. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5438–5448.

Yu, A., Ye, V., Tancik, M., Kanazawa, A., 2021. pixelnerf: Neural radiance fields from one or few images. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4578–4587.

Zhang, K., Riegler, G., Snavely, N., Koltun, V., 2020. NeRF++: Analyzing and Improving Neural Radiance Fields. *arXiv preprint arXiv:2010.07492*.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., Wang, O., 2018. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. *arXiv preprint arXiv:1801.03924*.