DLCV HW4

1

1a. the NeRF idea in your own words

NeRF represents a significant advancement in 3D modeling and rendering. In simple terms, it is a technique that utilizes deep learning to create highly detailed 3D models from a set of 2D images. The core idea is to represent a scene as a continuous 5D function (spatial location and viewing direction) using a neural network. When you input a spatial coordinate and a viewing angle into this model, it predicts the color and volume density at that point. By aggregating these predictions across different viewpoints, NeRF can generate novel, photorealistic images of the scene from perspectives not seen in the training images. This method is remarkable for its ability to capture complex light interactions like reflections and shadows, making it a powerful tool for creating immersive 3D environments.

1b. which part of NeRF do you think is the most important

The most crucial component of NeRF, in my opinion, is its ability to synthesize highly realistic images from sparse viewpoints. This is achieved through its unique architecture that models both color and density at each point in space, considering different viewing angles. This aspect is vital because it allows NeRF to interpolate and extrapolate scenes with great detail and accuracy, even in areas where direct image data is lacking. The network's training process, which involves adjusting the weights to minimize the difference between the observed and predicted images, is also instrumental in achieving high fidelity in image synthesis. This characteristic distinguishes NeRF from other 3D modeling techniques, making it particularly useful for applications in virtual reality, augmented reality, and visual effects.

1c. compare NeRF’s pros/cons w.r.t. other novel view synthesis work

NeRF has several advantages over traditional novel view synthesis methods. Its ability to produce highly realistic and detailed renders from a limited set of images is unparalleled. NeRF's renders exhibit remarkable consistency in lighting and material properties, outperforming earlier techniques in terms of realism and coherence. However, it's not without drawbacks. NeRF's computation is highly intensive, requiring significant processing power and time, which can be a limiting factor in practical applications. Additionally, NeRF struggles with dynamic scenes or changes in lighting, as it inherently assumes a static scene. In contrast, other methods like traditional 3D modeling or simpler deep learning approaches may offer faster rendering times or better adaptability to dynamic scenes but often at the cost of

lower image quality or realism. Therefore, while NeRF represents a major leap forward, it's best suited for applications where high fidelity and photorealism are more critical than computational efficiency or adaptability to changing scenes.

2

I mainly used the 2nd NeRF Github reference link provided by the TA, and only adjusted the `learning_rate` from 5e-4 to 1e-3, then compared the different settings as shown in Q3, and added a new `image_name` to the output of the dataset so that the generated image has the same name as the ground truth, and then manually moved the validation image out for evaluation, nothing else special.

3

Except for the Setting part of the form, it is the same.

number of coarse samples: 64

loss type: mse

batch size: 1024

chunk: 32*1024

optimizer: Adam

learning rate: 1e-3

scheduler: StepLR(gamma=0.1, step=20)

epoch: 1

Setting	PSNR	SSIM	LPIPS (vgg)
number of additional fine samples: 256 embedding_xyz: 12	*36.64*	*0.983*	*0.110*
number of additional fine samples: 256 embedding_xyz: 10	*35.92*	*0.980*	*0.123*
number of additional fine samples: 128 embedding_xyz: 10	*35.06*	*0.979*	*0.130*

Discussion:

Setting 1: (PSNR: 36.64, SSIM: 0.983, LPIPS: 0.110)

In the first setting with 256 additional fine samples and an `embedding_xyz` of 12, we observe the highest PSNR value of 36.64 which indicates that this setting achieves the best image reconstruction fidelity among the three settings. The SSIM score is very high at 0.983, suggesting that the perceptual quality of the image is also excellent, with structural details well-preserved. The LPIPS score, which measures perceptual similarity, is the lowest at 0.110, further corroborating that this setting results in images that are closer to the ground truth from a human perceptual standpoint.

Setting 2: (PSNR: 35.92, SSIM: 0.980, LPIPS: 0.123)

The second setting, with the same number of additional fine samples but a reduced `embedding_xyz` of 10, shows a slight decrease in performance. The PSNR drops to 35.92, indicating a minor reduction in reconstruction accuracy. This is mirrored by a slight decrease in SSIM to 0.980, suggesting that structural similarity to the ground truth has been affected, although not significantly. The LPIPS score increases to 0.123, implying that perceptually, this model produces images that are a bit less similar to the target compared to the first setting.

Setting 3: (PSNR: 35.06, SSIM: 0.979, LPIPS: 0.130)

The third setting with 128 additional fine samples and an `embedding_xyz` of 10 displays the lowest PSNR of 35.06. This suggests that halving the `number of fine samples` has a noticeable impact on the accuracy of image reconstruction. The SSIM also sees a slight decrement to 0.979, which might indicate that the reduced `number of fine samples` affects the preservation of structural information. The increase in LPIPS to 0.130, although minor, is consistent with the trend that fewer fine samples slightly degrade the perceptual quality of the image reconstruction.

In summary, there's a clear trend that shows increasing the `number of fine samples` and the `embedding_xyz` dimension improves the image quality across all metrics. The model is sensitive to these hyperparameters, and careful tuning is essential for optimal performance. It is worth noting that these results are based on a single epoch; more epochs might show different trends as the model continues to learn.

Definition:

PSNR: PSNR is a widely used metric in image processing for assessing the quality of reconstructed or compressed images compared to the original, high-quality

version. It is calculated using the mean squared error (MSE) between the two images. PSNR is expressed in decibels (dB), indicating the ratio of the maximum possible power of a signal to the power of corrupting noise. Higher PSNR values suggest better image quality. In the context of NeRF and deep learning, PSNR is crucial for evaluating the fidelity of synthesized 3D scenes or enhanced images.

SSIM: SSIM is a more sophisticated metric than PSNR for evaluating image quality. Instead of focusing solely on pixel-level differences, SSIM considers changes in structural information, luminance, and contrast. It evaluates the visual impact of three characteristics of an image: luminance, contrast, and structure, thereby providing a more holistic assessment of image quality. SSIM values range between -1 and 1, with higher values indicating greater similarity to the reference image. This metric is particularly useful in scenarios where preserving structural integrity is more important than exact pixel accuracy, like in certain deep learning image processing tasks.

LPIPS(vgg): LPIPS is a more recent and advanced metric for evaluating perceptual image quality. Unlike PSNR and SSIM, which are based on heuristically defined formulas, LPIPS uses a deep learning approach, often employing a pre-trained VGG network, to assess perceptual similarity. It measures the distance between two images in a feature space defined by the deep network. This approach aligns more closely with human visual perception, making it highly relevant for tasks in NeRF modeling where perceptual realism is crucial. LPIPS provides a more nuanced understanding of image quality, especially in contexts where traditional metrics like PSNR and SSIM may not align well with human perception. LPIPS values range between 0 and 1.

4