

CityNeRF: Building NeRF at City Scale

Yuanbo Xiangli^{1*}, Lining Xu^{1*}, Xingang Pan², Nanxuan Zhao^{1,3}, Anyi Rao¹, Christian Theobalt², Bo Dai⁴, Dahua Lin^{1,3,5}

¹The Chinese University of Hong Kong ²Max Planck Institute for Informatics ³CPII
⁴S-Lab, Nanyang Technological University ⁵Shanghai AI Laboratory

{xy019, xl020, ra018, dhlin}@ie.cuhk.edu.hk {xpan, theobalt}@mpi-inf.mpg.de
nanxuanzhao@gmail.com bo.dai@ntu.edu.sg

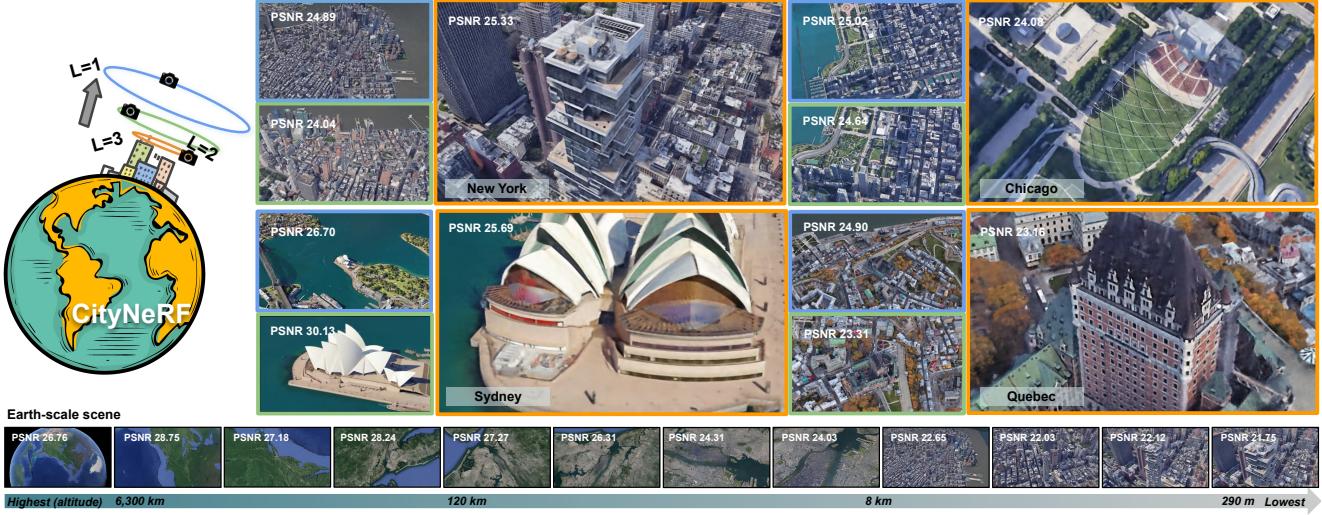


Figure 1. CityNeRF is capable of packing *city-scale* 3D scenes into a unified model, which preserves high-quality details across scales varying from satellite-level to ground-level. *Top*: We use the edge colors blue ($L=1$), green ($L=2$), and orange ($L=3$) to denote three scales from the most remote to the closest, with PSNR values displayed at the top-left corner of each rendered image. *Bottom*: CityNeRF can even accommodate variations at *earth-scale*. (src: New York, Chicago, Sydney, and Quebec scenes ©2021 Google)

Abstract

Neural Radiance Field (NeRF) has achieved outstanding performance in modeling 3D objects and controlled scenes, usually under a single scale. In this work, we make the first attempt to bring NeRF to city-scale, with views ranging from satellite-level that captures the overview of a city, to ground-level imagery showing complex details of an architecture. The wide span of camera distance to the scene yields multi-scale data with different levels of detail and spatial coverage, which poses great challenges to vanilla NeRF and biases it towards compromised results. To address these issues, we introduce CityNeRF, a progressive learning paradigm that grows the NeRF model and training set synchronously. Starting from fitting distant views with a shallow base block, as training progresses, new blocks are appended to accommodate the emerging details in the increasingly closer views. The strategy effectively activates high-frequency channels in the positional encoding and un-

folds more complex details as the training proceeds. We demonstrate the superiority of CityNeRF in modeling diverse city-scale scenes with drastically varying views, and its support for rendering views in different levels of detail. Project page can be found in [CityNeRF](#).

1. Introduction

Recently, neural volumetric representations [15, 16, 21, 23, 24] have demonstrated remarkable capability in learning to represent 3D objects and scenes from images, among which neural radiance field (NeRF) has been gaining increasing attention. NeRF [23] encodes a 3D scene with a continuous volumetric function parameterized by a multi-layer perceptron (MLP), and maps a 5D coordinate (position and viewing direction) to the corresponding color and volume density in a scene. While NeRF has been shown effective in controlled environments with a “bounded” and “single-scale” setting, it remains unclear if it is able to handle those scenarios that involve a wide range of scales.

However, the “single-scale” assumption can easily be violated in real-world scenarios. A typical case is our living city, which is essentially large-scale with adequate variety in components. Despite their popularity and versatility in various industries [3], city models are currently left out in the mainstream 3D scene datasets [8, 12, 22, 23, 36] for neural rendering, possibly due to their intrinsic complexity and diversity. In this work, we are interested in making the first attempt to build NeRF under *city-scale*.

A direct observation on city-scale scenes is that the camera is allowed a large degree of freedom in movement. Notably, the huge span in camera distance to the scene induces intrinsic *multi-scale*¹ characteristic: as the camera ascends, the ground objects in the scene are getting coarser appearances with less geometric detail and lower texture resolution; meanwhile, new objects from peripheral regions are getting streamingly included into the view with growing spatial coverage, as illustrated in Fig. 1. As the result, the variation in spatial coverage and levels of detail raises a tension among scales. It poses great challenges for vanilla NeRF, where the rendered remote views tend to be incomplete with artifacts in peripheral scene areas, and close views always have blurry textures and shapes.

Targeting the above challenges, we propose *CityNeRF* that adopts a multi-stage progressive learning paradigm. The whole training dataset is partitioned into a predefined number of scales according to the camera distances. Starting from the most remote scale, we gradually expand the training set by one closer scale at each training stage, and synchronously grow the model. In this way, CityNeRF robustly learns a hierarchy of representations across all scales of the scene. To facilitate the learning, two special designs are introduced: 1) *Growing model with residual block structure*: instead of naively deepening the MLP network, we grow the model by appending an additional block per training stage. Each block has its own output head that predicts the color and density residuals between successive stages, which encourages the block to focus on the emerging details in closer views; 2) *Inclusive multi-level data supervision*: the output head of each block is supervised by the union of images from the most remote scale up to its corresponding scale. In other words, the last block receives supervision from all training images while the earliest block is only exposed to the images of the coarsest scale. With such a design, each block module is able to fully utilize its capacity to model the increasingly complex details in closer views, and guarantees a consistent rendering quality between scales.

Extensive experiments show that our method constructs more complete remote views and brings out significantly more details in close views, as shown in Fig. 1, whereas vanilla NeRF and its derivatives constantly fail. Specifi-

cally, our method effectively preserves scene features learnt on remote views, and actively access higher frequency Fourier features in the positional encoding to construct finer details for close views. Furthermore, CityNeRF allows views to be rendered by different output heads from shallow to deep blocks, providing additional flexibility of viewing results in a *level-of-detail* manner.

2. Related Work

In this section, we first review a series of NeRF [23] derived works, focusing on their extended applicable scenarios and the role of position encoding. We further introduce the level-of-detail (LOD) concept that is related to multi-scale representations.

NeRF and beyond. NeRF has inspired many subsequent works that extend its continuous neural volumetric representation for more practical scenarios beyond simple static scenes, including unbounded scenes [37], dynamic scenes [13, 35], nonrigidly deforming objects [25–27, 33], phototourism settings with changing illumination and occluders [19], etc. The common adaptations include 1) conditioning NeRF on additional embedding (*e.g.*, instance appearance code [19]), and 2) representing with additional neural fields to learn extra features besides color and density for each query point (*e.g.*, warp field [25, 26], transient field [19]). In this work, we deal with a much wider scenario in terms of the spatial span, and aim to bring NeRF to an unprecedented *city-scale*. The above relaxations made on the static scene assumption are orthogonal to our efforts, and can be naturally integrated into CityNeRF to bring more advanced applications, such as modeling the time-of-day effects of a city with an additional appearance code [19].

Position Encoding. The adoption of position encoding in NeRF enables the multilayer perceptron (MLP) to learn high-frequency functions from the low-dimensional coordinate input. Tancik *et al.* [32] revealed the relationship between the positional encoding and the Neural Tangent Kernel (NTK), which pointed out that the Fourier feature mapping can transform the effective NTK into a stationary interpolating kernel with a tunable bandwidth. A small number of frequencies induces a wide kernel which causes underfitting of the data, while a large number of frequencies induces a narrow kernel causing over-fitting. A line of works have been proposed to improve NeRF by adjusting this position encoding input. Mip-NeRF [2] used integrated positional encoding (IPE) that replaces NeRF’s point casting with cone casting, which allows the model to reason about the 3D volumes. [14, 25, 26] alternatively adopt windowed positional encoding to aid training, where a parameter α is introduced to window the frequency band of the positional encoding.

Level-of-detail. In computer graphics, level-of-detail (LOD) [4, 17] refers to the complexity of a 3D model repre-

¹We refer interested readers to a short film *Cosmic Eye* (2012) to better understand the concept of *multi-scale* in nature.



Figure 2. Observed artifacts on vanilla NeRF [23] trained (a) jointly on all scales and (b) separately on each scale. Artifacts on rendered images are highlighted by solid boxes, with GT patches displayed on the side. (a) Joint training on all scales results in blurry texture in close views and incomplete geometry in remote views; (b) Separate training on each scale yields inconsistent geometries and textures between successive scales. (src: *New York* scene ©2021 Google)

sentation, in terms of metrics such as geometric detail and texture resolution. The hierarchical representation learnt by CityNeRF can be seen as an analogue to this concept, where features learnt at different blocks represent geometries and textures of different complexity. Recently, [31] proposed to represent implicit surfaces using an octree-based feature volume which adaptively fits shapes with multiple discrete LODs. [18] used a multi-scale block-coordinate decomposition approach that is optimized during training. All these approaches work under the assumption of a bounded scene, where the covered 3D volume/2D image can be divided into multi-granularity grids/blocks, similar to a quadtree or octree, whose vertices store scene features at different scales. Unlike these feature-based multi-scale representations, we take an alternative approach to represent multi-scale information with a dynamic model grown in blocks, which allows to represent scenes of arbitrary size, instead of relying on pre-learnt features on vertices within a bounded scene.

3. CityNeRF

In this paper, we aim at extending NeRF to city-scale scenes captured by cameras of very different distances. As illustrated in Fig. 1, the difference between the closest and the most remote camera can be thousands of kilometers away in real-world. This drastic span in camera distance brings multi-scale characteristic to city-scale scenes. The resulting levels of detail of the scene and the spatial coverage vary from one scale to another. Consequently, vanilla NeRF and its derivatives fail to capture peripheral scene contents or details of central objects as shown in Fig. 2.

In the following sections, Sec. 3.1 firstly briefs the necessary background for vanilla NeRF modeling. In Sec. 3.2, we discuss the challenges of *city-scale* scenes posed on vanilla NeRF. In Sec. 3.3, we explain our proposed progressive net-

work growing and training scheme.

3.1. Preliminaries on NeRF

NeRF [23] parameterizes the volumetric density and color as a function of input coordinates, using the weights of a multilayer perceptron (MLP). For each pixel displayed on the image, a ray $\mathbf{r}(t)$ is emitted from the camera's center of projection and passes through the pixel. A stratified sampling approach is proposed to determine a vector of sorted distances $\{t_k\}$ on the ray between the camera's predefined near and far planes. For any query point $\mathbf{r}(t_k)$ on the ray, the MLP takes in its Fourier transformed features, *i.e.* position encoding (PE), and outputs the color and density with:

$$(\tau_k, \mathbf{c}_k) = \text{MLP}(\gamma(\mathbf{r}(t_k))). \quad (1)$$

PE here is implemented by the concatenation of a set of sine and cosine mappings of the 3D position \mathbf{x} (and optionally viewing direction) up to a pre-defined frequency degree M :

$$\gamma(\mathbf{x}) = [\sin(2^0 \mathbf{x}), \cos(2^0 \mathbf{x}), \dots, \sin(2^{M-1} \mathbf{x}), \cos(2^{M-1} \mathbf{x})]^T, \quad (2)$$

where \mathbf{x} is normalized to lie in $[-\pi, \pi]$. The network is then optimized abiding by the classical volume rendering, where the estimated densities and colors for all the sampled points $\mathbf{r}(t_k)$ are used to approximate the volume rendering integral using numerical quadrature [20]:

$$\begin{aligned} \mathbf{C}(\mathbf{r}; \mathbf{t}) &= \sum_k T_k (1 - \exp(-\tau_k(t_{k+1} - t_k))) \mathbf{c}_k, \\ \text{with } T_k &= \exp\left(-\sum_{k' < k} \tau_{k'}(t_{k'+1} - t_{k'})\right), \end{aligned} \quad (3)$$

where $\mathbf{C}(\mathbf{r}; \mathbf{t})^2$ is the final predicted color of the pixel. The final loss is the total squared error between the rendered and

²We eliminate the notation for hierarchical sampling [23] for brevity.

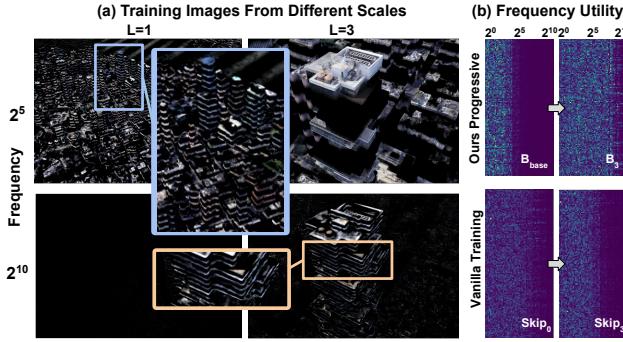


Figure 3. (a) Views of different scales require different Fourier feature frequencies to recover the details. While higher-frequency Fourier features (*e.g.* $\cos(2^{10}z)$) are necessary for close views, lower-frequency Fourier features (*e.g.* $\cos(2^5z)$) are sufficient to represent remote views. (b) Our progressive training strategy effectively activates the higher-frequency Fourier features provided in PE at deeper blocks (*e.g.* B_3), while vanilla NeRF constantly fails, even after the deepest skip connection.

ground truth colors:

$$\min_{\Theta} \sum_{\mathbf{r} \in \mathcal{R}} \left(\|\mathbf{C}^*(\mathbf{r}) - \mathbf{C}(\mathbf{r}; \mathbf{t})\|_2^2 \right), \quad (4)$$

where $\mathbf{C}^*(\mathbf{r})$ is the ground truth color, and \mathcal{R} is the collection of sampled rays within a batch.

3.2. Challenges on City-scale Scenes

The challenges brought by city-scale scenes are manifold. Firstly, the vastly different spatial coverage between remote and close views introduces a bias towards central scene contents, resulting in inconsistent quality within rendered images, as illustrated in Fig. 2(a). This is due to that central regions are repetitively captured by cameras across all scales, whereas peripheral regions are only observed by distant cameras. In contrast, training each scale separately eliminates such inconsistency but sacrifices the communication among different scales, leading to a significant discrepancy between successive scales, as shown in Fig. 2(b).

We also observe that the effective frequency channels in PE differ from one scale to another. As shown in Fig. 3(a), for a close view ($L = 3$) showing complex details of a rooftop, the low-frequency Fourier feature ($\cos(2^5z)$) appears to be insufficient, while a higher-frequency Fourier feature ($\cos(2^{10}z)$) is activated to better align with such details. In contrast, the remote view ($L = 1$) can be well represented by the low-frequency Fourier feature, hence the high-frequency one is damped. Subsequently, the high-frequency channels in PE are only activated in close views. However, due to the limited amount of close views in the training data, vanilla NeRF trained on all scales as a whole tend to overlook these high-frequency scene components, leading to compromised solutions that are biased to utilize low-frequency features only.

3.3. Progressive Model with Multi-level Supervision

The multi-scale characteristic of city-scale scenes implies different sample difficulty. We therefore propose to build and train the model in a progressive manner, with the aim to encourage the division of works among network layers, and unleash the power of the full frequency band in PE. Moreover, insights from curriculum learning [5–7, 9, 10, 30, 39] tell that by training models in a meaningful order, it may ease the training on difficult tasks with model weights initialized on easy samples. This further underpins our motivation to adopt a progressive training strategy. The overall paradigm of CityNeRF is shown in Fig. 4.

In our framework, the training data and the model are grown in a synchronized multi-stage fashion. We denote L_{max} as the total number of training stages pre-determined for a captured scene, which also discretizes the continuous distance between the camera and the scene. In our experiments, we chose to partition the training data according to the camera distance, which approximates a hierarchy of resolutions of objects in the scene, abide by projective geometry. We regard the closest view as at full resolution. A new scale is set when the camera zooms out by a factor of $\{2^1, 2^2, 2^3, \dots\}$.³ Each image I is then assigned with a stage indicator I_l that is shared among all its pixels and thus the sampling points along casted rays, with $\{I_l = L\}$ denoting the set of images belong to the scale L .

Start from remote views ($L = 1$), as the training progresses, views from one closer scale $L + 1$ are incorporated at each training stage. Such data feeding scheme remedies the bias in sample distribution by allowing the model to put more effort on peripheral regions at early training stage. Meanwhile, a rough scene layout can be constructed, which naturally serves as a foundation for closer views in subsequent training stages.

Along with the expansion of the training set, the model grows by appending new blocks. As shown in Fig. 4, each block is paired with an output head to predict the color and density *residuals* of scene contents viewed at successively closer scales. The most remote views are allowed to exit from the shallow block with only base color, while close-up views have to be processed by deeper blocks and rendered with progressively added residual colors. PE is injected to each block via a skip connection to capture the emerging complex details in scene components. All layers in the network remain trainable throughout the training process.

Residual Block Structure. It can be observed that remote views usually exhibit less complex details, making it a relatively easier task to start with. We adopt a shallow MLP to

³In general cases where the camera distances are not accessible, I_l can be approximated by the spatial size of textures in the image, or by investigating the frequency features using a trained vanilla NeRF similar to Fig 3(a). The choice of L_{max} is relative flexible, and can be reflected by the user's preference to desired division of levels of detail.

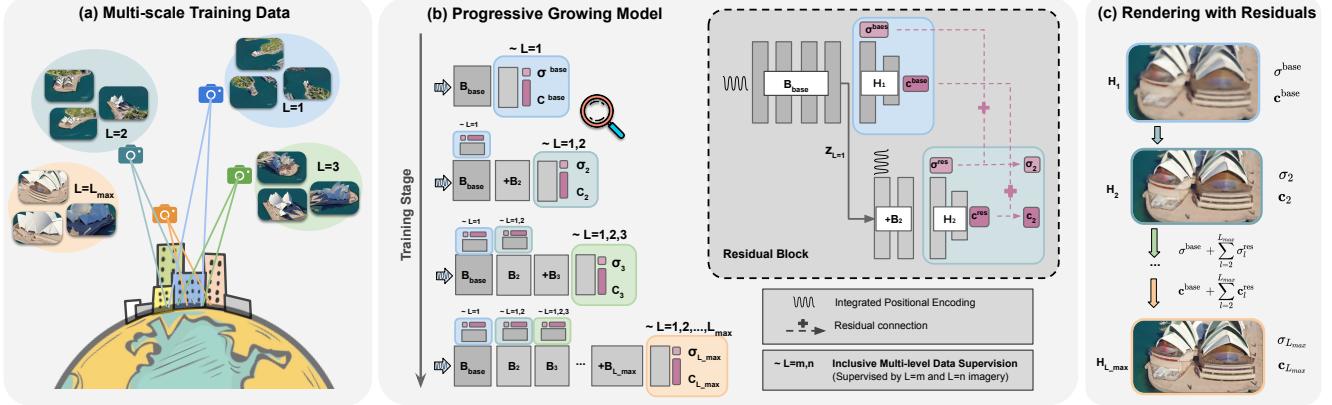


Figure 4. Overview of CityNeRF. (a) An illustration of the multi-scale data in city-scale scenes, where we use $L \in \{1, 2, 3, \dots\}$ to denote each scale. At each stage, our model grows in synchronization with the training set. (b) New residual blocks are appended to the network as the training proceeds, supervised by the union of samples from the most remote scale up to the current scale. The structure of a residual block is shown in the dashed box. (c) Level-of-detail rendering results obtained at different residual blocks. From shallow to deep, details are added bit by bit. (src: Sydney scene ©2021 Google)

be our *base block*, denoted as B_{base} , with $D_{base} = 4$ hidden layers and $W = 256$ hidden units each to fit the most remote scale $\{I_l = 1\}$. A skip connection is not included as the base block is shallow. The output head for color and density follows the original NeRF paper.

When we proceed to the next training stage, a block B_L consisting of $D_{res} = 2$ layers of non-linear mappings is appended to the model. A skip connection is added to forward the positional encoding $\gamma(\mathbf{x})$ to the block. The intuition is that, since shallow layers are fitted on remote views, the features are learnt to match with low level of detail, hence only low-frequency channels in PE are activated. However, the new layers need to access the high-frequency channels in PE to construct the emerging details in closer views. As verified in Fig. 3(b), our progressive training strategy is able to resort to higher-frequency Fourier features at a deeper block. In contrast, the matching baseline (*i.e.* Mip-NeRF-full) is incapable of activating high-frequency channels in PE even after the deepest skip layer, thus fails to represent more complex details.

The additive block B_L outputs residual colors and densities based on the latent features \mathbf{z}_{L-1} obtained from the last point transform layer of the previous training stage:

$$(\mathbf{c}_L^{res}, \sigma_L^{res}) = f_L^{res}(\mathbf{z}_{L-1}, \mathbf{x}, \mathbf{d}). \quad (5)$$

The output exit from head H_L is then aggregated as

$$\mathbf{c}_L = \mathbf{c}^{base} + \sum_{l=2}^L \mathbf{c}_l^{res}, \quad \sigma_L = \sigma^{base} + \sum_{l=2}^L \sigma_l^{res}. \quad (6)$$

The design with residuals has mutual benefits for all scales. Firstly, it encourages intermediate blocks to concentrate on the missing details and take advantage of the high-frequency Fourier features supplied via skip connection. Furthermore, it enables gradients obtained from latter blocks to smoothly

flow back to earlier blocks and enhance the shallow features with the supervision from closer views.

Inclusive Multi-level Supervision. To guarantee a consistent rendering quality across all scales, at training stage L , the output head H_L is supervised by the union of images from previous scales, *i.e.* $\{I_l \leq L\}$. The loss at stage L is aggregated over all previous output heads from H_1 to H_L :

$$\mathcal{L}_L = \sum_{l=1}^L \sum_{\mathbf{r} \in \mathcal{R}_l} \left(\left\| \hat{C}(\mathbf{r}) - C(\mathbf{r}) \right\|_2^2 \right), \quad (7)$$

where \mathcal{R}_l is the set of rays with stage indicator up to stage l , and $C(\mathbf{r}), \hat{C}(\mathbf{r})$ are the ground truth and predicted RGB.

The design of multi-level supervision embeds the idea of *level-of-detail*, with deeper output heads providing more complex details in the rendered views. Compared to traditional mipmapping [34] which requires a pyramid of pre-defined models at each scale, this strategy unifies different levels of detail into a single model and can be controlled with L .

4. Experiment

We train CityNeRF on multi-scale city data acquired from Google Earth Studio [1] and evaluate the quality of reconstructed views and synthesized novel views. We compare our method to NeRF [23], NeRF with windowed positional encoding (NeRF w/ WPE) [25], and Mip-NeRF [2]. The effects of the progressive strategy, and the block design with skip layer and residual connection are further analyzed in the ablation study.

Data Collection. Google Earth Studio [1] is used as the main data source for our experiments, considering their easy capturing of multi-scale city imagery by specifying camera altitudes. We test our model and compare against baselines on twelve city scenes across the world. When collecting

City Scene	Populated Cities (Main Experiments)		Places of Interest (Extra Tests: Used for various illustrations throughout the paper and additional comparisons in Supplementary.)											
	New York (56 Leonard)	San Francisco (Transamerica)	Sydney (Opera House)	Seattle (Space Needle)	Chicago (Pritzker Pavilion)	Quebec (Chateau Frontenac)	Amsterdam (New Church)	Barcelona (Sagrada Familia)	Rome (Colosseum)	Los Angeles (Hollywood)	Bilbao (Guggenheim)	Paris (Pompidou)		
Lowest Altitude (m)	290	326	115	271	365	166	95	299	130	660	163	159		
Highest Altitude (m)	3,389	2,962	2,136	18,091	6,511	3,390	2,3509	8,524	8,225	12,642	7,260	2,710		
Number of Images	463	455					220							

Table 1. Twelve city scenes captured in Google Earth Studio.

	56 Leonard (PSNR ↑)				Transamerica (PSNR ↑)				56 Leonard (Avg.)			Transamerica (Avg.)		
	Stage I.	Stage II.	Stage III.	Stage IV.	Stage I.	Stage II.	Stage III.	Stage IV.	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑
NeRF (D=8, Skip=4)	21.279	22.053	22.100	21.853	22.711	22.811	22.976	21.581	21.702	0.320	0.636	22.642	0.318	0.690
NeRF w/ WPE (D=8, Skip=4)	22.022	22.097	21.799	21.439	23.382	23.171	22.450	20.806	21.672	0.365	0.633	22.352	0.331	0.680
Mip-NeRF-small (D=8, Skip=4)	21.899	22.179	22.045	21.763	23.346	23.030	22.645	20.937	21.975	0.344	0.648	22.692	0.327	0.687
Mip-NeRF-large (D=10, Skip=4)	22.020	22.403	22.277	21.975	23.196	22.900	22.439	20.743	22.227	0.318	0.666	22.525	0.330	0.686
Mip-NeRF-full (D=10, Skip=4,6,8)	22.043	22.484	22.690	22.355	23.377	23.156	22.854	21.152	22.312	0.266	0.689	22.828	0.314	0.699
CityNeRF	23.145	23.548	23.744	23.015	24.259	23.911	23.507	22.942	23.481	0.235	0.739	23.606	0.265	0.749
CityNeRF (Longer Time)	24.120	24.345	25.382	25.112	24.608	24.350	24.357	24.608	24.513	0.160	0.815	24.415	0.192	0.801

Table 2. Quantitative comparison on *56 Leonard* (*New York*) and *Transamerica Pyramid* (*San Francisco*) scenes. D means model depth and *Skip* indicates which layer(s) the skip connection is inserted to. Better performance can be achieved if each stage is trained for a longer time e.g. until convergence. (**best** / 2nd best)

data, we move the camera in a circular motion and gradually elevate the camera from a low altitude (\sim ground-level) to a high altitude (\sim satellite-level). The radius of the orbit trajectory is expanded during camera ascent to ensure a large enough spatial coverage. Statistics of the collected scenes are listed in Tab. 1.

Metrics. We report all metrics on the results exit from the last output head $H_{L_{max}}$ in CityNeRF. For quantitative comparison, results are evaluated on low-level full reference metrics, including the Peak Signal-to-Noise Ratio (PSNR) which computes MSE in log space, and Structural Similarity Index Measure (SSIM) [29] that reflects human visual perception. Perceptually, we report LPIPS [38] metric, which computes MSE between normalized features from all layers of a pre-trained VGG encoder [28]. Since the scenes are captured across different scales, a single mean metric averaged from all rendering results cannot fairly reflect quality. For example, an image taken at satellite-level with less complex details in shape and texture can easily obtain a higher PSNR than the closer ones captured near the ground. We thus additionally report the mean PSNR obtained at each scale for fairness.

Implementation. We set $L_{max} = 4$ for CityNeRF, hence the model at the final training stage has 10 layers of 256 hidden units, with skip connections at the 4, 6, 8-th layer. For fair comparison, the best baseline model with the same configuration is also implemented. We adopt Mip-NeRF’s cone-based integrated positional encoding [2], as we found it consistently outperforms NeRF’s position encoding. For all methods, the highest frequency is set to 2^{10} , with 128 sample queries per ray. We train CityNeRF for 100k iterations per stage and baselines till converge (approx. 300-400k iterations). All models are optimized using Adam [11] with a learning rate decayed exponentially from $5e^{-4}$ and a batch size of 2,048. The optimizer is reset at each training stage for CityNeRF.

4.1. Experiment Results

Tab. 2 shows our experiment results obtained at the final training phase on two populated city scenes. Extras

scenes listed in Tab. 2 serve as various illustrations throughout the paper, with qualitative comparisons provided in Supplementary. In Fig. 5 we display the rendered novel views using CityNeRF and baseline methods on *New York scene*.

We show that naively deepening the network cannot resolve the problems arisen from different levels of detail and spatial coverage among scales, where remote views still suffer from blurry artifacts at edges. On the other hand, CityNeRF attains a superior visual quality both on the entire dataset and at each scale on all metrics, with a notable improvement in remote scales, where peripheral areas in the rendered views appear to be clearer and more complete compared to jointly training on all images. When approaching the central target as the camera descends, CityNeRF continuously brings more details to the scene components, whereas baseline methods always result in blurry background objects. At the finest scale, CityNeRF yields the most detailed central target and achieves a decent visual quality in background objects that is on par with the view center. Fig. 3(b) visualizes the network weights associated to each channel of PE, where a clear shift towards the higher frequency portion indicates that the model is able to leverage these part of the information to construct details in the view. A more detailed weights evolution plot can be found in supplementary.

Fig. 6 qualitatively shows the rendering results from different output heads. It can be noticed that H_2 produces the coarsest visual results, which omits a significant amount of details when zooming in to closer views, but appears to be plausible for the set of remote views. The latter output heads gradually add more complex geometric and texture details to the coarse output from previous stage, while maintaining the features learnt at shallower layers meaningful to earlier output heads. In practice, one may consider using earlier heads for rendering remote views for the sake of the storage and time consumption.

4.2. Ablation Study

Effectiveness of Progressive Strategy. The effectiveness of our progressive learning strategy is analyzed through

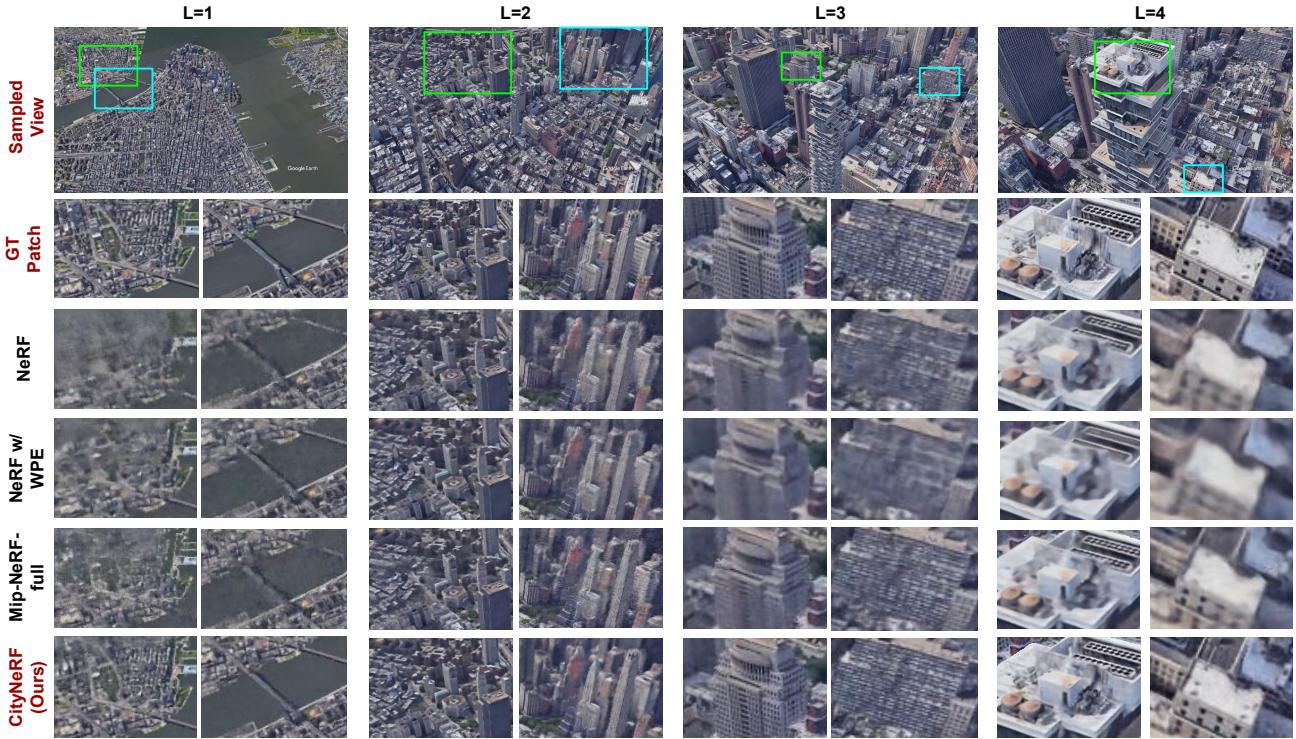


Figure 5. Qualitative comparisons between NeRF [23], NeRF w/ WPE [25], Mip-NeRF-full [2], and CityNeRF. CityNeRF consistently outperforms baseline methods across all scales with reliably constructed details. (src: New York scene ©2021 Google)

	56 Leonard (PSNR ↑)				Transamerica (PSNR ↑)				56 Leonard (Avg.)			Transamerica (Avg.)		
	Stage I.	Stage II.	Stage III.	Stage IV.	Stage I.	Stage II.	Stage III.	Stage IV.	PSNR ↑	LPIPS ↓	SSIM ↑	PSNR ↑	LPIPS ↓	SSIM ↑
w/o inclusive multi-level sup.	22.570	22.475	23.300	23.188	23.623	23.039	22.508	21.294	22.773	0.298	0.700	22.789	0.309	0.696
w/o data feeding schedule	22.268	22.660	22.380	21.949	23.563	23.355	23.015	21.366	22.340	0.324	0.679	23.015	0.310	0.711
w/o model growing	22.826	22.121	21.471	20.972	24.026	23.560	22.554	20.503	22.139	0.326	0.666	23.055	0.317	0.706
w/o skip layer	23.046	23.364	23.317	22.746	24.232	23.696	22.950	22.151	23.159	0.259	0.730	23.235	0.283	0.725
w/o residual	23.104	23.325	23.228	22.786	24.091	23.608	23.236	22.796	23.139	0.257	0.729	23.364	0.272	0.735
Ours Complete Model	23.145	23.548	23.744	23.015	24.259	23.911	23.507	22.942	23.481	0.235	0.739	23.606	0.265	0.749

Table 3. Ablation studies on 56 Leonard (New York) and Transamerica Pyramid (San Francisco) scene.

three ablation studies: 1) Ablate the inclusive multi-level data supervision, *i.e.* replacing the output head supervision $\{I_l \leq L\}$ with $\{I_l = L\}$; 2) Ablate the progressive data feeding schedule, *i.e.* train on all scales simultaneously, with different scales predicted by different output heads. The model is fixed with $L = 4$ throughout the training; 3) Ablate model progression and only use the output head H_4 of the final block. Note we still adopt the progressive data feeding schedule for this case. Results are listed in Tab. 3.

It is observed that: 1) Without the inclusive multi-level data supervision, the model achieves a slightly higher PSNR at the closest scale ($L = 4$), but the performances at remote scales degrade drastically. This is because deeper blocks are solely trained on close views, and are not responsible for constructing remote views. As the result, the additive blocks might deviate to better fit the close views, whilst shallow layers just maintain their status and ignore the extra information from deeper layers. 2) Without the progressive data feeding strategy, the results are even inferior than baselines. First of all, the problem of imbalanced sample distribution remains unsolved, hence remote views are still poorly modeled. Due to the residual connection, the subop-

timal results set a less ideal foundation for the modeling of closer views. Secondly, with all data being fitted simultaneously, the effective channels in PE still can not be distinguished between scales. On top of this, regularizing shallow features with remote views puts more restrictions on model capacity, which further harms the performance. 3) Without appending new layers, it is difficult for the model to accommodate the newly involved high-frequency information from closer scales, where the model has already been well fitted on distant scales. As the result, it achieves decent performance on the most remote scale ($L = 1$) but becomes worse at closer scales.

Effectiveness of Model Design. To determine the effectiveness of the residual connection and the skip connection in the block structure, we run ablations on: 1) Discarding the residual connection when growing the network; 2) Inserting a skip connection at the same position as the original NeRF and discarding the rest in the additive blocks.

From Tab. 3, we can see that without the skip layer or the residual connection, performance at each scale slightly degenerates but still outperform baselines. 1) Qualitatively, it can be observed that, the influence of skip layer is more

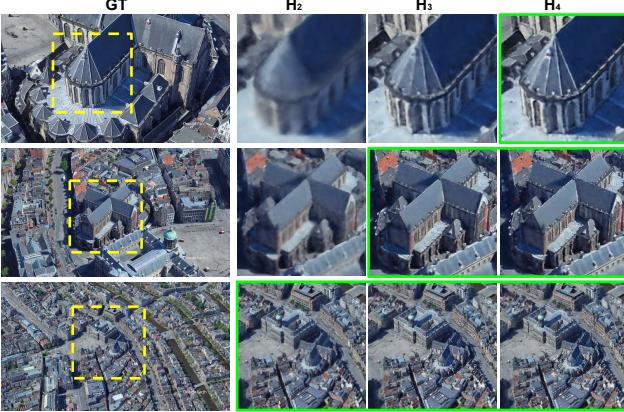


Figure 6. Rendering results obtained from different output heads. CityNeRF allows flexible exits from different residual blocks with controllable levels of detail. Note that remote views can exit from earlier heads with sufficient image details, and close views can get finer details when exiting from latter blocks. src: *Amsterdam* scene ©2021 Google)

notable on closer scales. In particular, one can see that on *Transamerica* scene, the absence of skip connection severely harms the performance at the closest scale $L = 4$, which can be ascribed to the fact that its remote views are much easier than close views. 2) Residual connections help form a better scene feature at remote scales for shallow output heads, by leveraging the supervision from deeper blocks to help correct the errors in early training phases. Consequently, rendering results of lower LOD are more accurate, which in turn benefit latter training phases on closer views. Furthermore, we found that as the training evolves (e.g., from $L = 2$ to $L = 4$), CityNeRF with residual gains an increasing advantage compared to the counterpart without residual. The additive nature of residual densities and colors enforces the model to emphasize the errors between the images rendered by previous stages and the ground truth, guiding the deeper blocks to recover the missing details using the newly introduced frequency information. Meanwhile, it also improves the predictions from previous output heads, yielding better base geometries and colors. More examples can be found in supplementary.

4.3. Extensions

To test the generalizability of CityNeRF, extra experiments running on real-world drone data (Fig. 8), earth-scale scenes (Fig. 1), and multi-dive trajectories have been conducted. CityNeRF successfully handles all these cases with verified robustness and high quality. We refer interested readers to supplementary for more details.

5. Conclusion

In this work, we propose CityNeRF that enables NeRF to model city-scale scenes. Specifically, we identify the challenges brought by the intrinsic multi-scale character-

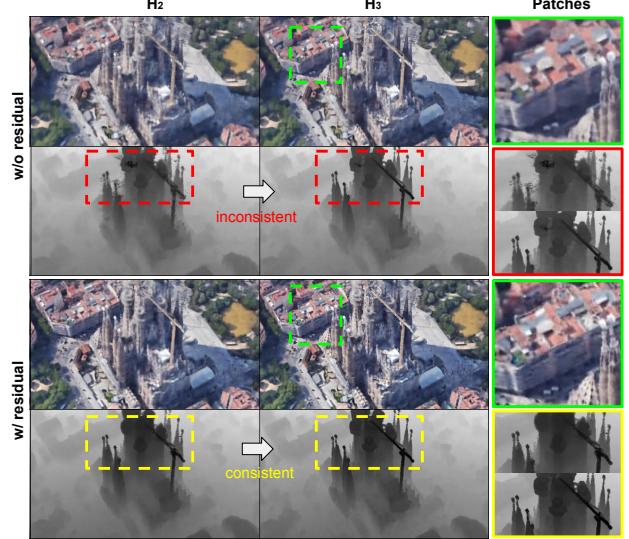


Figure 7. The residual connection enables supervisions from latter block heads to help refine the outputs from earlier heads, yielding more accurate geometries with consistent depths across scales. It also helps CityNeRF to focus on the missing details between the results rendered by shallower blocks and ground truth, leading to sharper visuals. (src: *Barcelona* scene ©2021 Google)

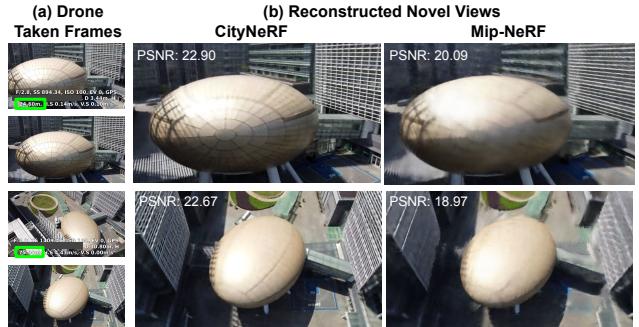


Figure 8. Results on real-world drone scenes. (a) Video frames with recorded camera information show the wide range of camera altitude changes (e.g., 24~76m in the two frames presented here). (b) Novel views rendered with CityNeRF trained in a simplified two-scale setting demonstrate the superior performance compared to Mip-NeRF [2].

istic, where a vanilla NeRF has difficulty in accommodating different levels of detail and spatial coverage between scales. Targeting this issue, we adopt a progressive training paradigm that synchronously grows the model and training set to learn a hierarchy of scene representations from coarse to fine. Our approach achieves superior results on city-scale scenes compared to multiple baselines, with distant views being more complete and close views being more delicate. CityNeRF sets a foundation for modeling drastically larger 3D scenes in real-world, and exploits the potential of neural radiance fields to be utilized in industrial applications.

References

- [1] Google earth studio. <https://earth.google.com/studio/>. 5
- [2] Jonathan T Barron, Ben Mildenhall, Matthew Tancik, Peter Hedman, Ricardo Martin-Brualla, and Pratul P Srinivasan. Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. *arXiv preprint arXiv:2103.13415*, 2021. 2, 5, 6, 7, 8
- [3] Filip Biljecki, Jantien Stoter, Hugo Ledoux, Sisi Zlatanova, and Arzu Cöltekin. Applications of 3d city models: State of the art review. *ISPRS International Journal of Geo-Information*, 4(4):2842–2889, 2015. 2
- [4] James H Clark. Hierarchical geometric models for visible surface algorithms. *Communications of the ACM*, 19(10):547–554, 1976. 2
- [5] Xiyang Dai, Dongdong Chen, Mengchen Liu, Yinpeng Chen, and Lu Yuan. Da-nas: Data adapted pruning for efficient neural architecture search. *ArXiv*, abs/2003.12563, 2020. 4
- [6] Sheng Guo, Weilin Huang, Haozhi Zhang, Chenfan Zhuang, Dengke Dong, Matthew R. Scott, and Dinglong Huang. Curriculumnet: Weakly supervised learning from large-scale web images. *ArXiv*, abs/1808.01097, 2018. 4
- [7] Yong Guo, Yaofu Chen, Yin Zheng, Peilin Zhao, Jian Chen, Junzhou Huang, and Mingkui Tan. Breaking the curse of space explosion: Towards efficient nas with curriculum search. *ArXiv*, abs/2007.07197, 2020. 4
- [8] Rasmus Jensen, Anders Dahl, George Vogiatzis, Engin Tola, and Henrik Aanæs. Large scale multi-view stereopsis evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 406–413, 2014. 2
- [9] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 4
- [10] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8107–8116, 2020. 4
- [11] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6
- [12] Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics (ToG)*, 36(4):1–13, 2017. 2
- [13] Zhengqi Li, Simon Niklaus, Noah Snavely, and Oliver Wang. Neural scene flow fields for space-time view synthesis of dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6498–6508, 2021. 2
- [14] Chen-Hsuan Lin, Wei-Chiu Ma, Antonio Torralba, and Simon Lucey. Barf: Bundle-adjusting neural radiance fields. *arXiv preprint arXiv:2104.06405*, 2021. 2
- [15] Lingjie Liu, Jatao Gu, Kyaw Zaw Lin, Tat-Seng Chua, and Christian Theobalt. Neural sparse voxel fields. *ArXiv*, abs/2007.11571, 2020. 1
- [16] Stephen Lombardi, Tomas Simon, Jason M. Saragih, Gabriel Schwartz, Andreas M. Lehrmann, and Yaser Sheikh. Neural volumes. *ACM Transactions on Graphics (TOG)*, 38:1 – 14, 2019. 1
- [17] David Luebke, Martin Reddy, Jonathan D Cohen, Amitabh Varshney, Benjamin Watson, and Robert Huebner. *Level of detail for 3D graphics*. Morgan Kaufmann, 2003. 2
- [18] Julien NP Martel, David B Lindell, Connor Z Lin, Eric R Chan, Marco Monteiro, and Gordon Wetzstein. Acorn: Adaptive coordinate networks for neural scene representation. *arXiv preprint arXiv:2105.02788*, 2021. 3
- [19] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7210–7219, 2021. 2
- [20] Nelson Max. Optical models for direct volume rendering. *IEEE TVCG*, 1995. 3
- [21] Lars M. Mescheder, Michael Oechsle, Michael Niemeyer, Sebastian Nowozin, and Andreas Geiger. Occupancy networks: Learning 3d reconstruction in function space. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4455–4465, 2019. 1
- [22] Ben Mildenhall, Pratul P Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 38(4):1–14, 2019. 2
- [23] 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 *European conference on computer vision*, pages 405–421. Springer, 2020. 1, 2, 3, 5, 7
- [24] Jeong Joon Park, Peter R. Florence, Julian Straub, Richard A. Newcombe, and S. Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 165–174, 2019. 1
- [25] Keunhong Park, Utkarsh Sinha, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Steven M Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5865–5874, 2021. 2, 5, 7
- [26] Keunhong Park, Utkarsh Sinha, Peter Hedman, Jonathan T Barron, Sofien Bouaziz, Dan B Goldman, Ricardo Martin-Brualla, and Steven M Seitz. Hypernerf: A higher-dimensional representation for topologically varying neural radiance fields. *arXiv preprint arXiv:2106.13228*, 2021. 2
- [27] Albert Pumarola, Enric Corona, Gerard Pons-Moll, and Francesc Moreno-Noguer. D-nerf: Neural radiance fields for dynamic scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10318–10327, 2021. 2
- [28] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015. 6

- [29] Vincent Sitzmann, Michael Zollhoefer, and Gordon Wetzstein. Scene representation networks: Continuous 3d-structure-aware neural scene representations. *ArXiv*, abs/1906.01618, 2019. 6
- [30] Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. Curriculum learning: A survey. *arXiv preprint arXiv:2101.10382*, 2021. 4
- [31] Towaki Takikawa, Joey Litalien, Kangxue Yin, Karsten Kreis, Charles Loop, Derek Nowrouzezahrai, Alec Jacobson, Morgan McGuire, and Sanja Fidler. Neural geometric level of detail: Real-time rendering with implicit 3d shapes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11358–11367, 2021. 3
- [32] Matthew Tancik, Pratul P Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *arXiv preprint arXiv:2006.10739*, 2020. 2
- [33] Edgar Tretschk, Ayush Tewari, Vladislav Golyanik, Michael Zollhofer, Christoph Lassner, and Christian Theobalt. Non-rigid neural radiance fields: Reconstruction and novel view synthesis of a dynamic scene from monocular video. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12959–12970, 2021. 2
- [34] Lance Williams. Pyramidal parametrics. In *Proceedings of the 10th annual conference on Computer graphics and interactive techniques*, pages 1–11, 1983. 5
- [35] Wenqi Xian, Jia-Bin Huang, Johannes Kopf, and Changil Kim. Space-time neural irradiance fields for free-viewpoint video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9421–9431, 2021. 2
- [36] Yao Yao, Zixin Luo, Shiwei Li, Jingyang Zhang, Yufan Ren, Lei Zhou, Tian Fang, and Long Quan. Blendedmvs: A large-scale dataset for generalized multi-view stereo networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1790–1799, 2020. 2
- [37] Kai Zhang, Gernot Riegler, Noah Snavely, and Vladlen Koltun. Nerf++: Analyzing and improving neural radiance fields. *arXiv preprint arXiv:2010.07492*, 2020. 2
- [38] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 6
- [39] Tianyi Zhou, Shengjie Wang, and Jeff A. Bilmes. Robust curriculum learning: from clean label detection to noisy label self-correction. In *ICLR*, 2021. 4