Supplementary Material for
In-Place Scene Labelling and Understanding with Implicit Scene Representation

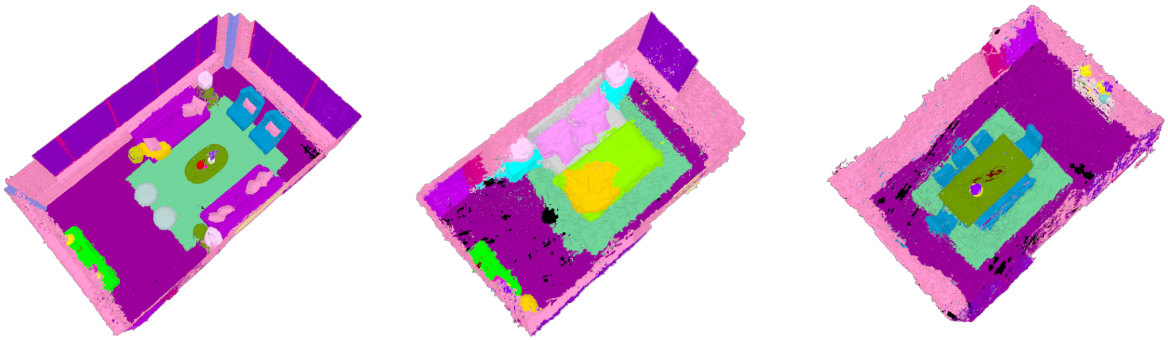

Figure 1: Semantic 3D reconstruction obtained using Semantic-NeRF. Note that our learned scene-specific 3D representation predicts decent geometry and semantics in occluded regions and fills the holes caused by unobserved regions to some extent.

2D Depth Metrics	
Abs Rel	$\frac{1}{n}\sum \lvert d - d_{gt}\rvert / d_{gt}$
Abs Diff	$\frac{1}{n}\sum \lvert d - d_{gt}\rvert$
Sq Rel	$\frac{1}{n}\sum \lvert d - d_{gt}\rvert^2 / d_{gt}$
RMSE	$\sqrt{\frac{1}{n}\sum \lvert d - d_{gt}\rvert^2}$
$\delta < 1.25^i$	$\frac{1}{n}\sum (\max(\frac{d}{d_{gt}}, \frac{d_{gt}}{d}) < 1.25^i)$

Table 1: Definitions of depth metrics: n is the number of valid depth pixels, d and d_{gt} are rendered depths at testing poses and ground truth depths, respectively.

A. Effects of Learning Semantics to Radiance and Geometry

Corresponding to Section 4.2 of the main paper, Table 2 shows quantitative results for photometric and geometric reconstruction quality when projected to 2D on Replica scenes with and without semantics enabled. We observe no obvious difference between these two set-ups. Peak signal-to-noise ratio (PSNR) is used to measure the quality of the rendered colour images and the metrics used to evaluate the 2D depth maps are shown in Table 1.

B. Semantic 3D Reconstruction from Posed Images

After training semantic-NeRF with in-place annotation, we can also extract an explicit 3D scene from the learned MLP to inspect the implicit 3D representation. Geometric meshes are extracted by first querying the MLP on dense 3D grids of the scene and then applying marching cubes. Attached semantic texture is rendered by treating the *negative* normal direction of vertices in the mesh as the ray marching directions during volume rendering. We show qualitative results for three Replica room scenes in Figure 1.

C. Network Architecture

Axis-aligned positional encoding (PE) of 3D positions are fed to both first and intermediate fully-connected (FC) layers with 256 neurons and ReLU activations before predicting volume density. Additional FC layers with 128 neurons are used for *view-invariant* semantics and *view-dependent* radiance after merging input viewing directions.

The length of positional encoding L relates to the maximum frequency used and affects the rendering quality. In label propagation task, we find that using only low-frequency components ($L = 5$) leads to over-smoothed 2D renderings, while using high-frequency ones ($L = 40$) leads to noisy interpolations, which aligns with findings in recent literature [16, 30]. L of 10 empirically performs the best.

D. More Qualitative Results

Here we show more examples of qualitative results in Figure 2, 3, 4 for semantic view synthesis, label denoising and super-resolution, respectively.

We kindly urge readers to watch our *supplementary video* on project page https://shuaifengzhi.com/Semantic-NeRF/ which highlights the accuracy and consistency of semantic renderings in various situations and applications.

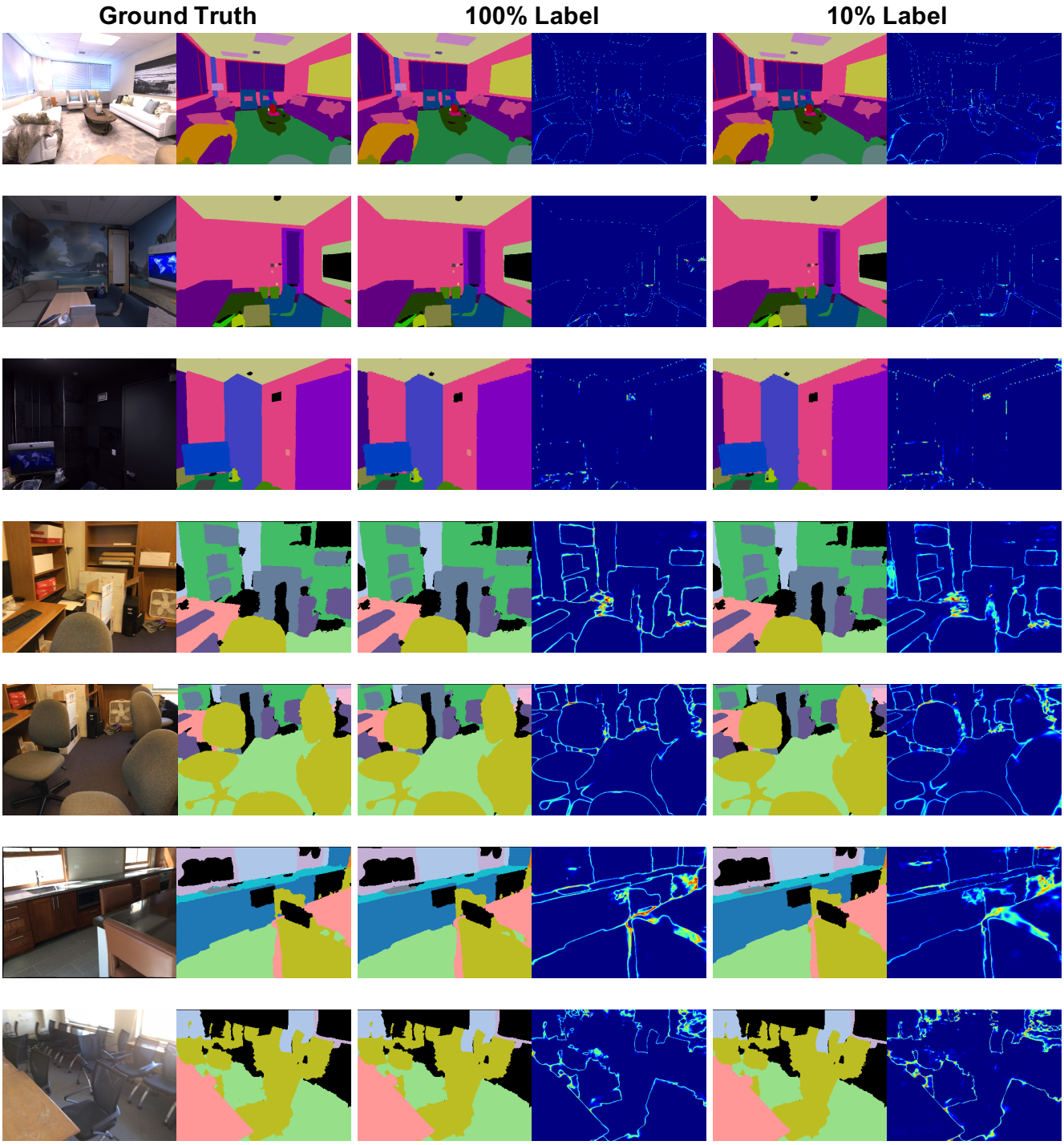

Figure 2: View-synthesis results.

Network Set-up	Depth							RGB
	AbsRel↓	AbsDiff↓	SqRel↓	RMSE↓	$\delta < 1.25$↑	$\delta < (1.25)^2$↑	$\delta < (1.25)^3$↑	PSNR↑
W/ Semantics	0.017	0.032	0.007	0.096	0.993	0.997	0.998	32.27
W/O Semantics	0.018	0.032	0.009	0.102	0.993	0.996	0.998	32.80

Table 2: Quantitative evaluation of effects of predicting semantics on appearance and geometry on Replica dataset.

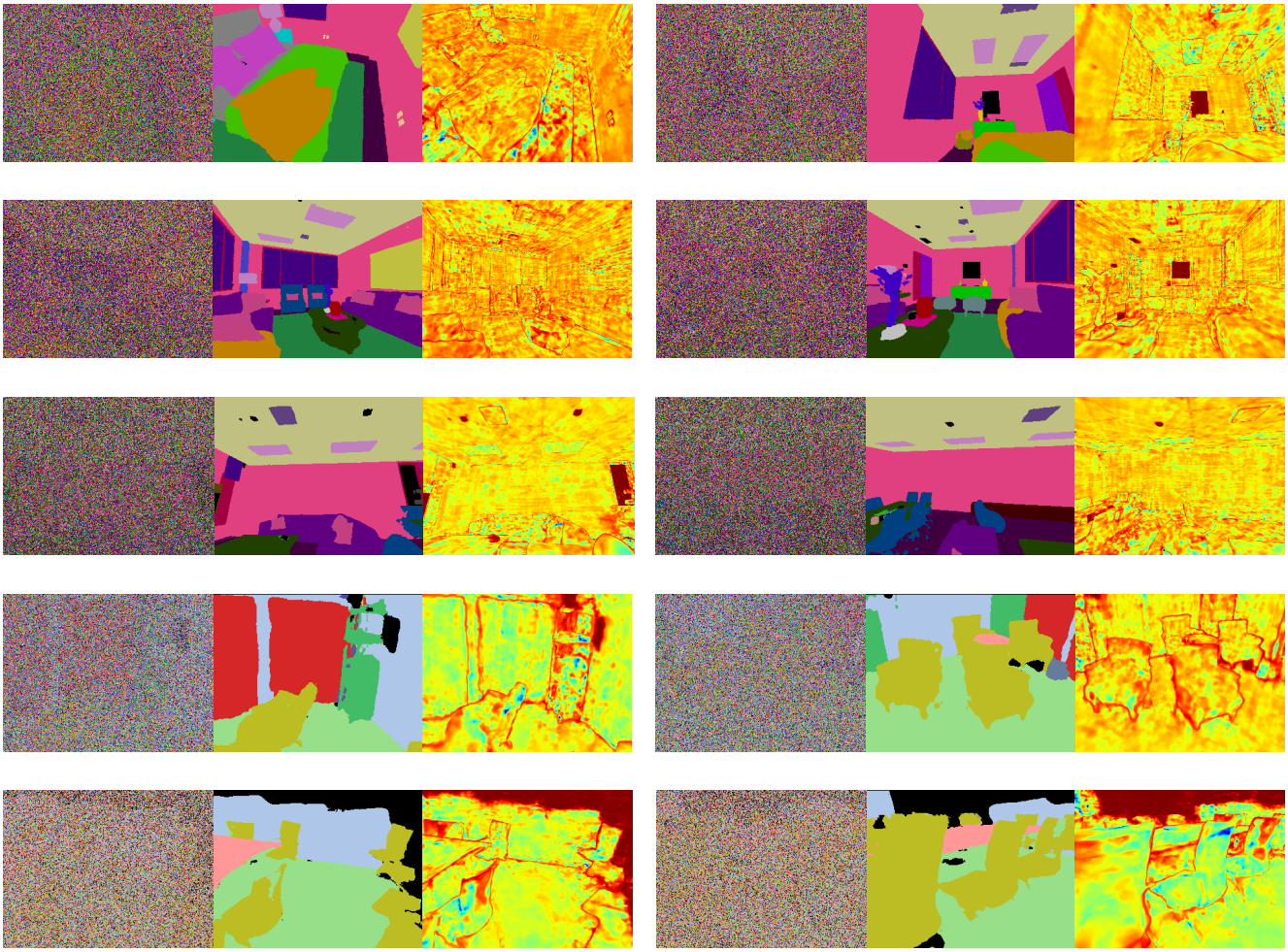

Figure 3: Pixel-wise denoising of semantic labels with 90% noise ratio.

(a) Super-resolution using coarse label

(b) Super-resolution using sparse label

Figure 4: Label super-resolution (×8) results.