RGBD Based Dimensional Decomposition Residual Network for 3D Semantic Scene Completion – Supplementary Material

1. More Details of the Proposed Network

In Table 1, we show the details of a dilated DDR block. DDR (k, w, s, d) is used to denote the DDR block with the kernel size k, the output channels of feature maps w, the stride s and the dilation rate d. PWConv represents the point-wise convolution and DDRConv represents the proposed DDR convolution within a DDR block. The details of the proposed network structure are shown in Table 2. EWAdd represents the element-wise add.

Operation	Kernel	Channels	Stride	Dilation
PWConv	$1 \times 1 \times 1$	w/4	1	1
DDRConv	$1 \times 1 \times k$	w/4	s	d
DDRConv	$1 \times k \times 1$	w/4	s	d
DDRConv	$k \times 1 \times 1$	w/4	s	d
PWConv	$1 \times 1 \times 1$	w	1	1

Table 1. Details of the DDR(k, w, s, d). k is the kernel size, w is the output channels of the feature map, s is the stride and d is the dilation rate.

2. More Qualitative Results

2.1. Results of Using Two-modality Information as Input

Figure 1 list some visualized results on NYUCAD. As can be seen in Figure 1(1)-(3), the proposed light-weight SSC network can achieve better results than SSCNet, and is much more accurate in shape completion and semantic segmentation.

When a object is composed of several parts with various textures, the color information may cause confusion for semantic labeling locally, *e.g.* in Figure 1(4), one part of the wall is red and another part is white, and the segmentation for the adjacent area of two colors has some errors. However, the semantic labeling still benefits from the color information as a whole.

2.2. Results of Using Single-modality information as Input

In Figure 2, we list the results by using depth or RGB image as input. For some cases, when objects can not be distinguished solely rely on the depth information, color image can provide quite useful cues for the labeling purpose. For instance, in Figure 2(1)-(4), the predictions using color image as inputs are more accurate than the one that use depth as input. Although our method has been proved to be capable of in-cooperating depth and color information, it unexpectedly fail for few cases. For example, in Figure 2(5), the semantic labeling performance is not good even with color information. This may be due to the fact that there is an imbalance between the depth and color information when they are merged, and the network is dominated by only one of the two modalities in this case.

Module	Operation	Output Size	Kernel	Stride	Dilation
Feature Extractor	PWConv_1	$640 \times 480 \times 8$	1	1	1
	DDR_2d_1	$640 \times 480 \times 8$	3	1	1
	DDR_2d_2	$640 \times 480 \times 8$	3	1	1
	Projection layer	$240 \times 144 \times 240 \times 8$	-	-	-
	MaxPooling_1	$120 \times 72 \times 120 \times 8$	2	2	1
	Conv_1	$120 \times 72 \times 120 \times 8$	3	2	1
	Concatenate	$120 \times 72 \times 120 \times 16$	[MaxPooling_1, Conv_1]		
	DDR_3d_1	$120 \times 72 \times 120 \times 16$	3	1	1
	MaxPooling_2	$60 \times 36 \times 60 \times 16$	2	2	1
	Conv_2	$60 \times 36 \times 60 \times 48$	3	2	1
	Concatenate	$60 \times 36 \times 60 \times 64$	[MaxPooling_2, Conv_2]		
	DDR_3d_2	$60 \times 36 \times 60 \times 64$	3	1	1
Feature Fusion	EWAdd_1	$60 \times 36 \times 60 \times 64$	-	-	-
	DDR_3d_3	$60 \times 36 \times 60 \times 64$	3	1	2
	EWAdd_2	$60 \times 36 \times 60 \times 64$	-	-	-
	DDR_3d_4	$60 \times 36 \times 60 \times 64$	3	1	3
	EWAdd_3	$60 \times 36 \times 60 \times 64$	-	-	-
	DDR_3d_5	$60 \times 36 \times 60 \times 64$	3	1	5
	EWAdd_4	$60 \times 36 \times 60 \times 64$	-	-	-
	Concatenate	$60 \times 36 \times 60 \times 256$	[EWAdd_1, EWAdd_2,EWAdd_3,EWAdd_4]		
LW-ASPP	PWConv_2	$60 \times 36 \times 60 \times 64$	1	1	1
	DDR_3d_6	$60 \times 36 \times 60 \times 64$	3	1	3
	DDR_3d_7	$60 \times 36 \times 60 \times 64$	3	1	6
	DDR_3d_8	$60 \times 36 \times 60 \times 64$	3	1	9
	GlobalAvgPool	$60 \times 36 \times 60 \times 64$	-	-	-
	Concatenate	$60 \times 36 \times 60 \times 320$	[PWConv_2,DDR_3d_6,DDR_3d_7,DDR_3d_8,GlobalAvgPool]		
Output	PWConv	$60 \times 36 \times 60 \times 128$	1	1	1
	PWConv	$60 \times 36 \times 60 \times 128$	1	1	1
	PWConv	$60 \times 36 \times 60 \times 12$	1	1	1
	ArgMax	$60 \times 36 \times 60 \times 12$	-	-	-

Table 2. The details of the proposed network architecture.

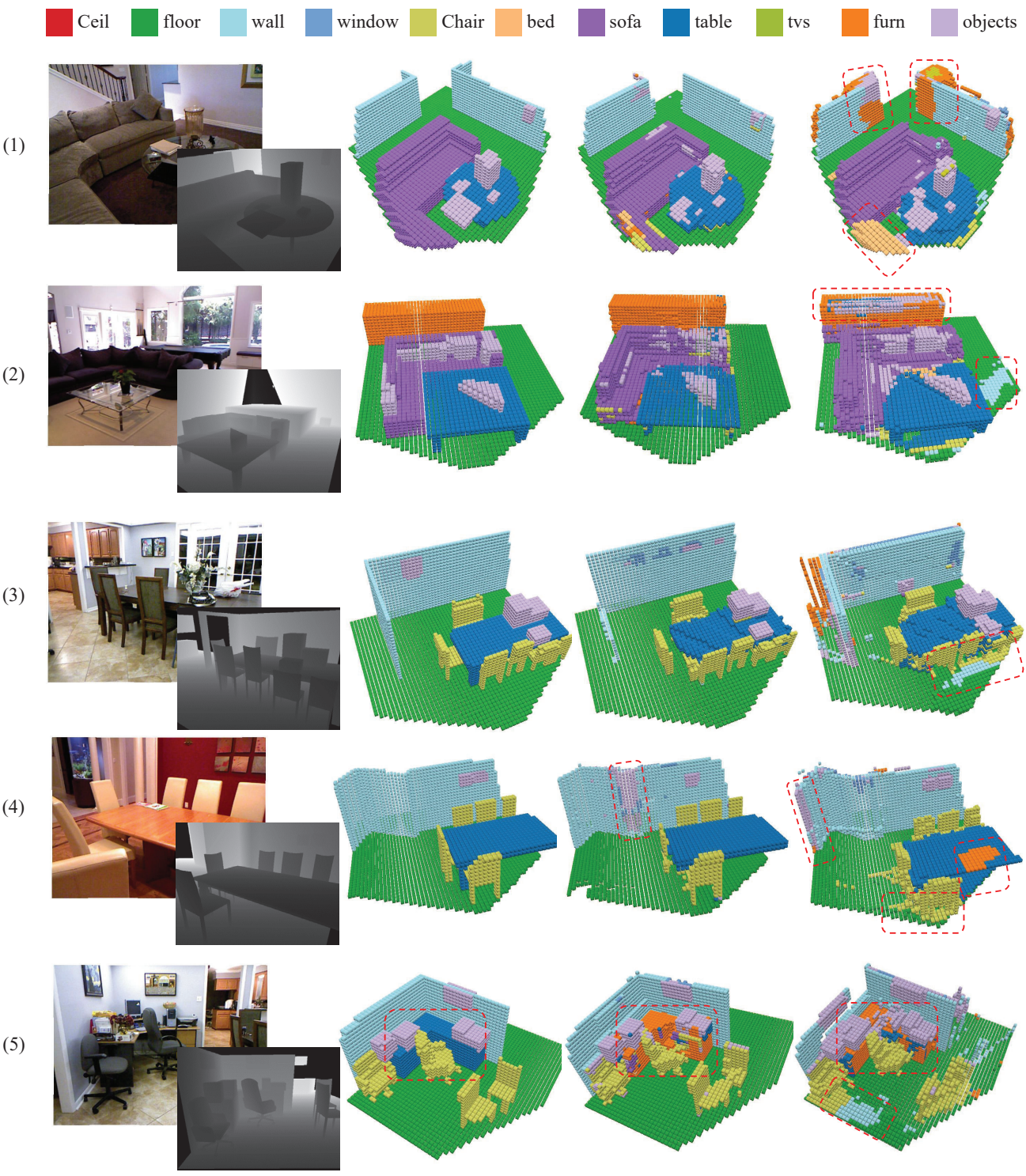

Figure 1. Qualitative results on NYUCAD. From left to right: Input RGB-D images, ground truth, results obtained by the proposed method, and results obtained by SSCNet. Overall, our completed semantic 3D scenes are less cluttered and much accurate compared to SSCNet.

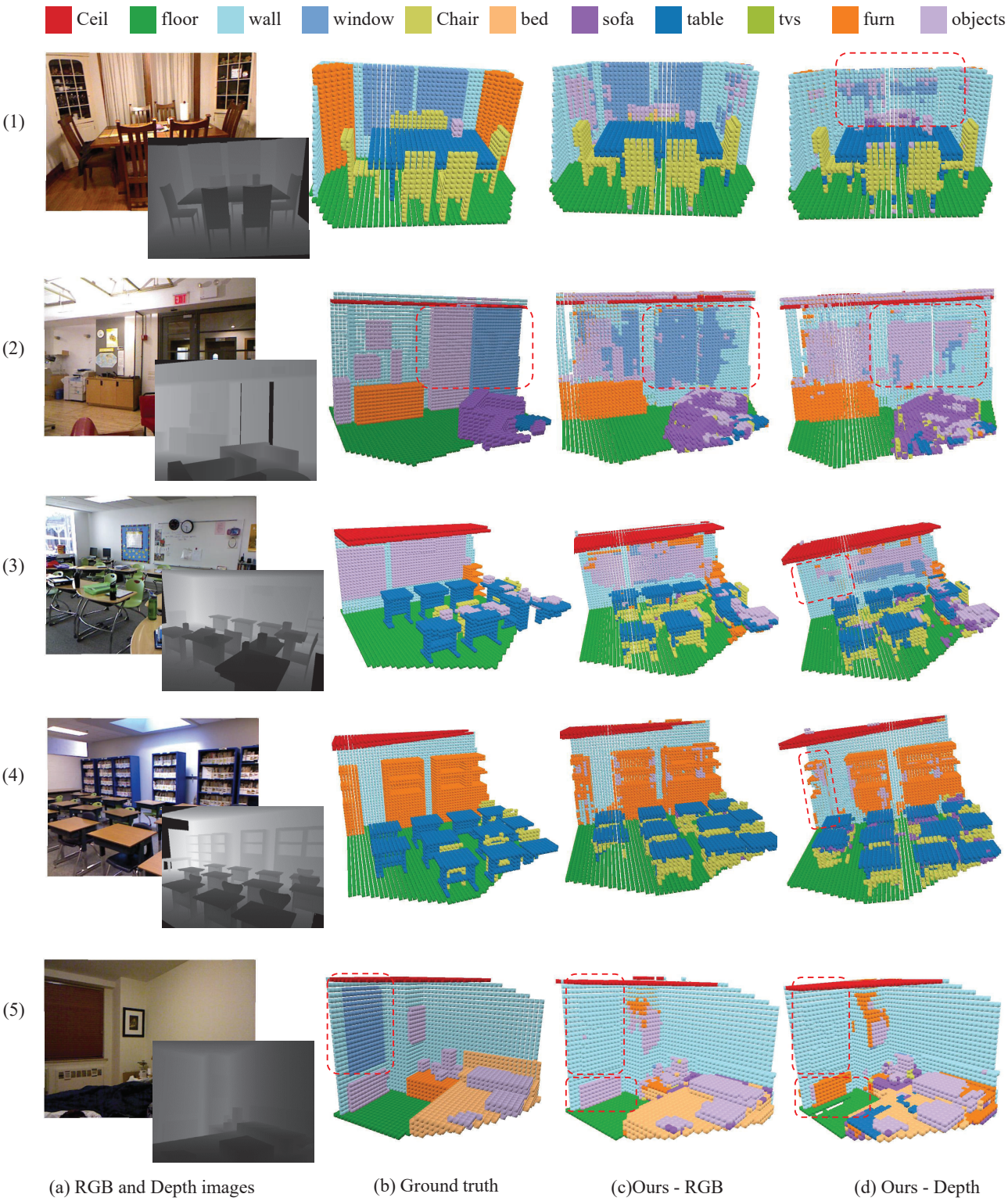

Figure 2. Results of using sole Depth or RGB image as input.