RT-GS2: Real-Time Generalizable Semantic Segmentation for 3D Gaussian Representations of Radiance Fields: Supplementary Material

Mihnea-Bogdan Jurca*[1, 2]
mihnea-bogdan.jurca@vub.be
Remco Royen*[1]
remco.royen@vub.be
Ion Giosan[2]
ion.giosan@cs.utcluj.ro
Adrian Munteanu[1]
adrian.munteanu@vub.be

[1] Department ETRO
Vrije Universiteit Brussel
Brussels, Belgium

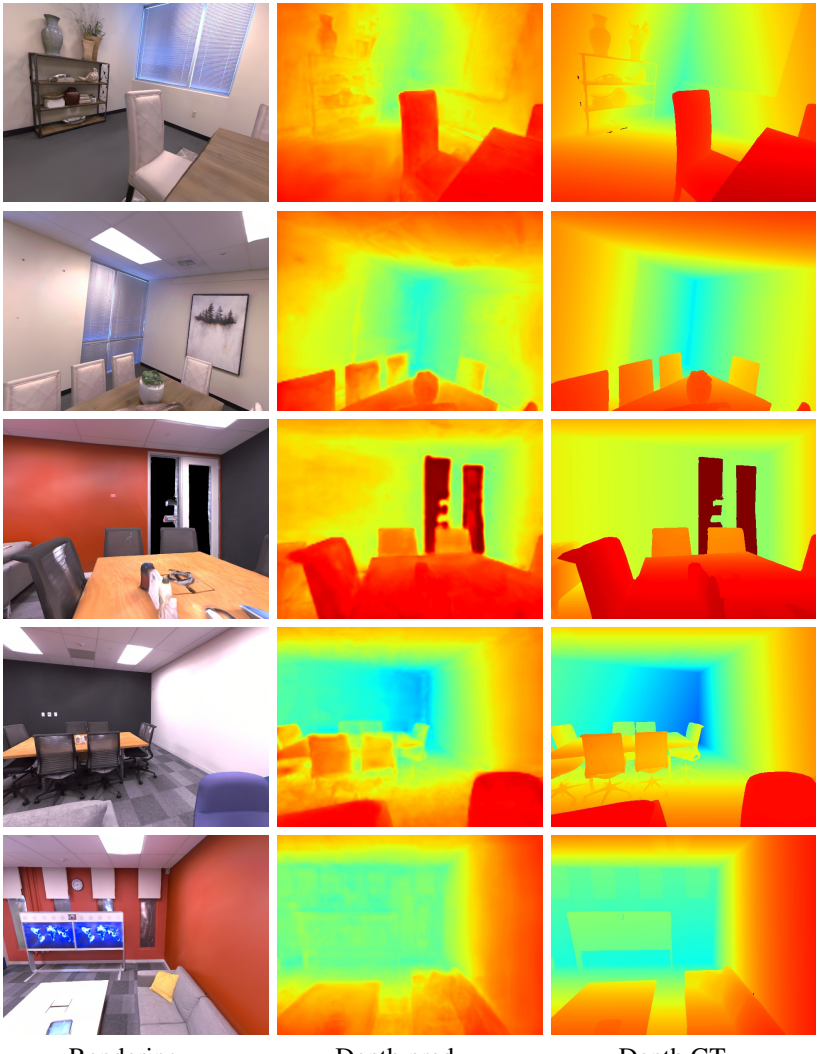
[2] Computer Science Department
Technical University of Cluj-Napoca,
Cluj-Napoca, Romania

1 Robustness of self-supervised features.

In order to evaluate the robustness of the self-supervised view-independent 3D Gaussian features we utilize them for another downstream task, namely depth prediction. Instead of doing monocular depth prediction on the rendered image, we employ our splatted 3D Gaussian features as additional information. While we did not re-train the 3D Gaussian feature learning, we trained another VDVI feature fusion module and depth prediction head. A schematic overview of the employed architecture can be found in Figure 2 of the main body. To evaluate our results quantitatively, we employed the popular metrics, employed in [7]. In Table 1, we describe the different employed metrics, where d and d^* are the true and predicted depth values of a pixel, respectively, and n the total number of pixels in the instance. To train the VDVI feature fusion module and depth prediction head, we employed the MSE-loss as loss function.

The results of the proposed method and ablation of the 3D Gaussian features on the Replica dataset [8] can be found in Table 2. It can be seen that the addition of our view-independent 3D Gaussian features, allows a consistent improvement of the depth prediction for all employed metrics. More specifically, we achieve an important 24.1% improvement in Abs. Rel. and 25.4% in RMSE, compared to the experiment without self-supervised view-independent 3D Gaussian features, i.e. monocular depth prediction. Qualitative results are presented in Figure 1, visualized using a heatmap going from red, closeby, to blue, far away. It can be noticed that the depth predictions are of high quality, both closeby and far away, closely resembling the ground-truth depth maps.

* Both authors contributed equally to the paper.



Rendering Depth pred. Depth GT

Figure 1: Depth prediction of the proposed method on Replica dataset.

Abs Rel	$\frac{1}{n}\sum\frac{\lvert d-d^*\rvert}{d^*}$
Abs Diff	$\frac{1}{n}\sum\lvert d-d^*\rvert$
Sq Rel	$\frac{1}{n}\sum\frac{\lvert d-d^*\rvert^2}{d^*}$
RMSE	$\sqrt{\frac{1}{n}\sum\lvert d-d^*\rvert^2}$
$\delta < 1.25^i$	$\frac{1}{n}\sum\left(\max\left(\frac{d}{d^*},\frac{d^*}{d}\right) < 1.25^i\right)$
Comp	% valid predictions

Table 1: Definition of the employed depth metrics [7]

	Abs Rel	Abs Diff	Sq Rel	RMSE	δ<1.25	δ<1.25^2	δ<1.25^3	Comp
Ours w/o 3D features	0.108	0.238	0.040	0.303	0.911	0.985	0.994	0.998
Ours	**0.082**	**0.173**	**0.024**	**0.226**	**0.948**	**0.994**	**0.998**	**0.999**

Table 2: Performance metrics on Replica dataset [8] and ablation of the effect of the self-supervised view-independent 3D Gaussian features for depth prediction. W/o stands for without.

2 Details of the loss functions

2.1 Contrastive loss

Our robust view-independent 3D features are learnt in a self-supervised manner, using the contrastive point cloud loss described in [10]. Additionaly, specifically for operating on 3D Gaussian representations, we added scale and opacity transformations to the training input to ensure that the existing information of the Gaussians is retained during training. We will follow below the notations from the main body of our paper, for consistency reasons and facilitate reader comprehension.

Given a set of k scenes with their corresponding $\mathbf{G}^k$ representation, we construct a list of correspondences $P_1,\ldots,P_k$. For two different viewpoints of a scene k, we define the set $\{\mathbf{g}_1^k\prime,\mathbf{g}_2^k\prime,\ldots\}$ as the points that lie within the frustum of both views. Indices m and n denote the positions of the points in the first and second views, respectively, as listed in P_k. These correspondences are considered positive pairs and retain their positive value for contrastive loss computation. The employed loss-function can be expressed as follows:

$$L_{cl} = -\sum_{(m,n)\in P_k}\log\frac{\exp(f_m^k\cdot f_n^k/\tau)}{\sum_{(l,\cdot)\in P_k}\exp(f_m^k\cdot f_l^k/\tau)}, \tag{1}$$

where τ is a constant set to 0.07.

2.2 Semantic loss

The employed semantic loss function is composed by two terms: the per pixel cross entropy loss and the *CeCo* term, expressed mathematically as follows

$$\mathcal{L}_{sem} = \mathcal{L}_{CrossEntropy} + \lambda_{CeCo}\mathcal{L}_{CeCo}. \tag{2}$$

The *CeCo* loss-term, $\mathcal{L}_{CeCo}$, described in [13], can be expressed mathematically as follows:

$$\mathcal{L}_{CeCo}(\bar{Z}, \mathbf{W}^*) = -\sum_{k=1}^{K} \log\left(\frac{\exp(\bar{z}_k^\top \mathbf{w}_k^*)}{\sum_{k'=1}^{K} \exp(\bar{z}_{k'}^\top \mathbf{w}_{k'}^*)}\right), \tag{3}$$

Where $\bar{Z}, \mathbf{W}^*$ are the features centers and the classifier weights, respectively. In this scenario K describes the number of classes. The intuition of using this addition term was due to the highly unbalanced nature of the 2D segmentation labels, as classes such as walls, ceiling and floors are dominant. This intuition was supported by the results of our ablation study, in Table 4 of the main body, where can be seen that *CeCo* proves mostly beneficial for the less dominant classes. This can be noticed as the performance increase for the experiment with all classes is larger than the performance increase for only 20 most frequent classes.

3 Dataset setup

In this section, we discuss our dataset setup in more detail. Our experiments were conducted on three different datasets: Replica [8], ScanNet [2], and ScanNet++ [11]. For the Replica split, we followed the Semantic-NeRF setup as described in [12]. The tested resolution was 480×640. For the ScanNet dataset, we trained on the first 60 scenes and tested on 10 scenes, following the setup in [1]. The tested resolution was the same as in [1, 5]. For ScanNet++, we randomly selected 40 scenes for training and 10 scenes for testing.

4 More implementation details

4.1 View-independent implementation details

For the self-supervised view-independent 3D Gaussian feature learning, we used PointTransformerV3 [9] as the encoder, optimized with the Adam optimizer [4] ($\beta_1 = 0.9$, $\beta_2 = 0.999$) and a weight decay of 10^{-5}. The number of points queried for contrastive learning is 4096. To select the different views for scenes, we ensure that the corresponding frustums for those views have an overlap of at least 30% but no more than 80%.

4.2 View-Dependent / View-Independent (VDVI) feature fusion implementation details

We used the backbone of Asymformer [3], extracting the last activations before the final layer to provide the appropriate information for L_{cl}. Optimization was performed using the AdamW optimizer [6] ($\beta_1 = 0.9$, $\beta_2 = 0.999$) with a weight decay of 10^{-4}. During the generalization stage, we applied a warm-up of 4 epochs, which was disregarded during the fine-tuning stage. The learning rate was set to 10^{-4}.

5 Video Demo

Attached to the supplementary materials is also a video demo in which we exhibit the results for different video sequences, displaying the extraordinary performance of RT-GS2 over a complete sequence. Sequences were selected for both a synthetic dataset, Replica [8], and

real-world data, ScanNet++ [11], exhibiting the robustness of our highly accurate results. It can be noticed from the video demo, that the proposed method enhances view-consistency for our selected downstream tasks.

6 Additional visualizations

In this section, we show additional qualitative visualizations on all three datasets. Figure 2 and Figure 3 contain additional visualizations from Replica [8] and ScanNet [2], respectively. Figure 4 and Figure 5 present extensive visualizations on ScanNet++ [11] for all 10 test scenes on which experiments were conducted.

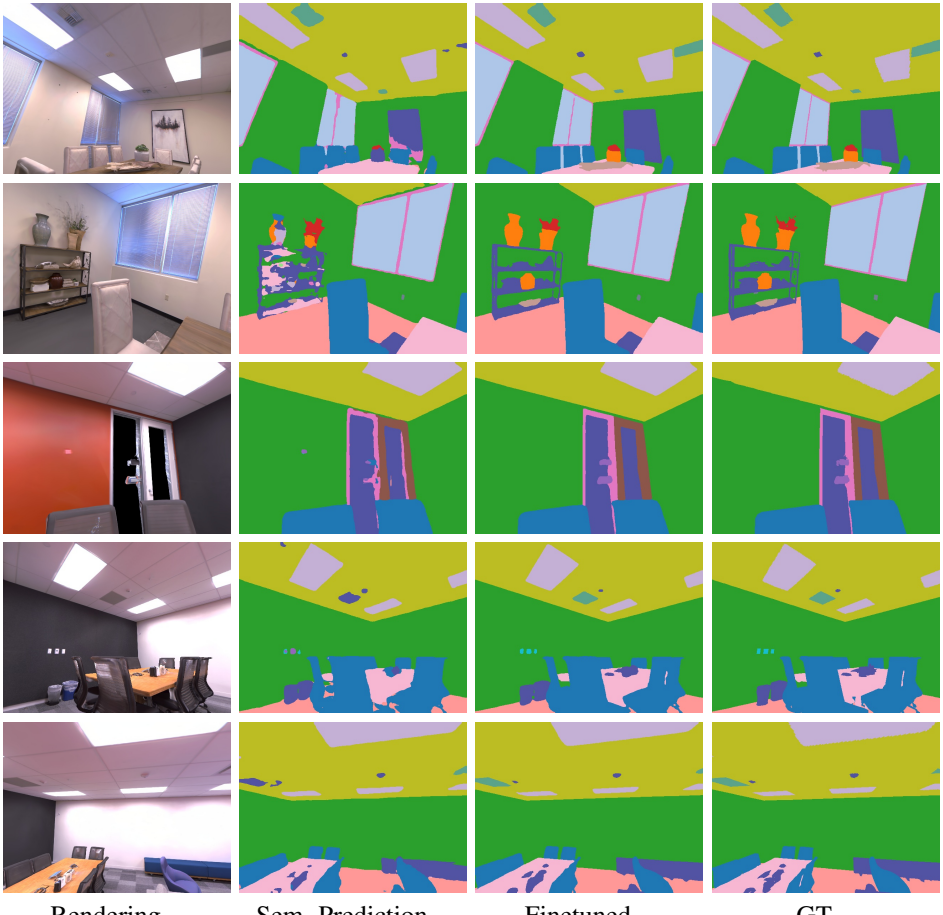


Figure 2: Additional qualitative results of RT-GS2 on the Replica [8] dataset.

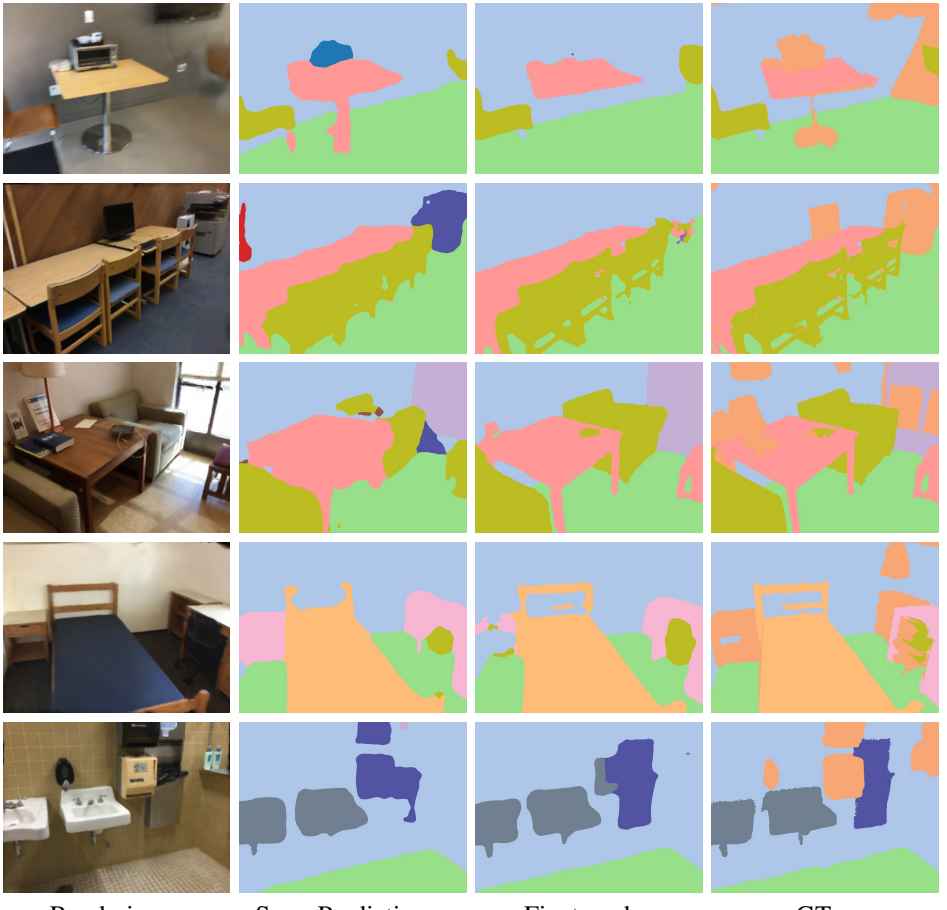
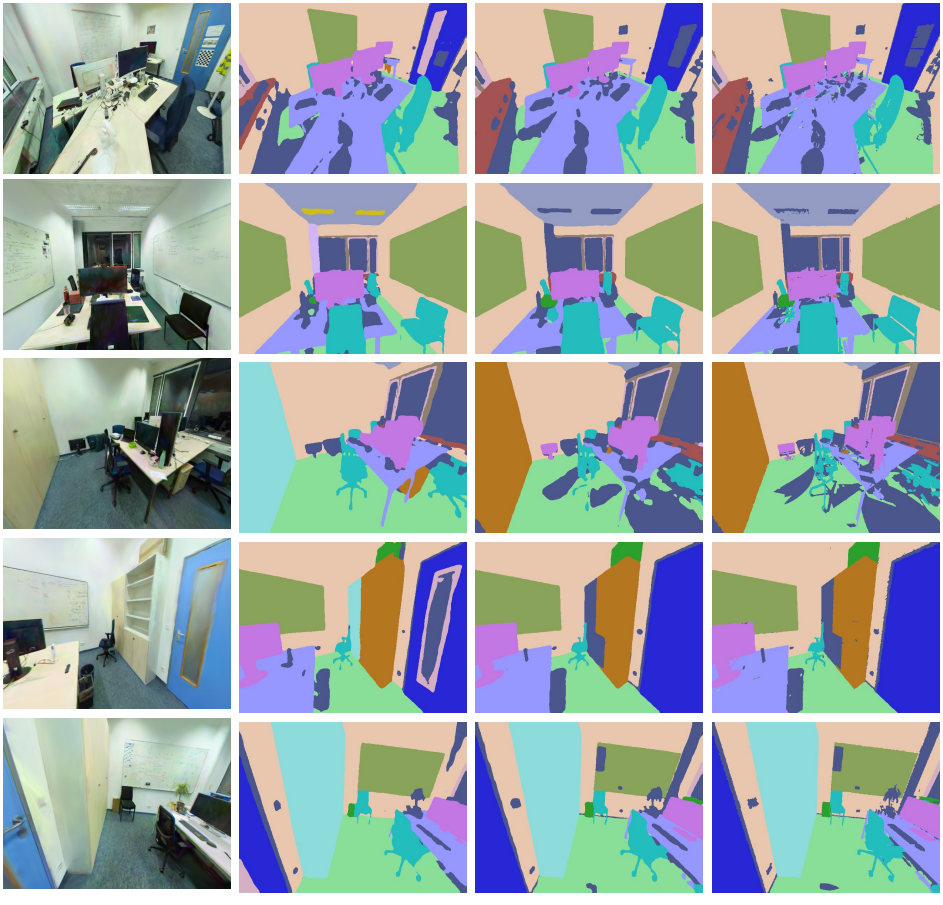
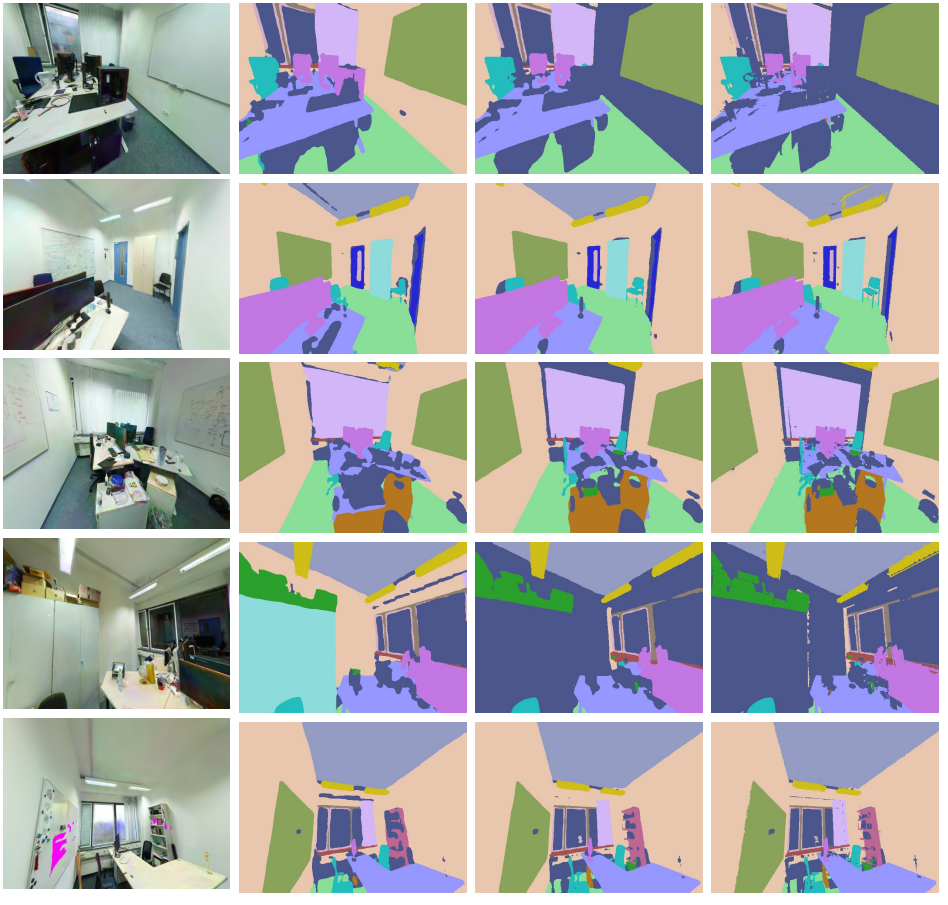


Figure 3: Additional qualitative results of RT-GS2 on the ScanNet [2] dataset. We point out that the peach orange color is the unannotated class, which is frequently present in the ScanNet dataset.



Rendering Prediction Finetuned GT

Figure 4: Qualitative results on ScanNet++ [] on the first five test scenes. We point out that dark purple color represents the unannotated and other classes in the ScanNet++ dataset.



Rendering Prediction Finetuned GT

Figure 5: Qualitative results on ScanNet++ [] on the last five test scenes. We point out that dark purple color represents the unannotated and other classes in the ScanNet++ dataset.

References

[1] Zi-Ting Chou, Sheng-Yu Huang, I Liu, Yu-Chiang Frank Wang, et al. Gsnerf: Generalizable semantic neural radiance fields with enhanced 3d scene understanding. *arXiv preprint arXiv:2403.03608*, 2024.

[2] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[3] Siqi Du, Weixi Wang, Renzhong Guo, and Shengjun Tang. Asymformer: Asymmetrical cross-modal representation learning for mobile platform real-time rgb-d semantic segmentation. *arXiv preprint arXiv:2309.14065*, 2023.

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[5] Fangfu Liu, Chubin Zhang, Yu Zheng, and Yueqi Duan. Semantic ray: Learning a generalizable semantic field with cross-reprojection attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17386–17396, 2023.

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[7] Zak Murez, Tarrence Van As, James Bartolozzi, Ayan Sinha, Vijay Badrinarayanan, and Andrew Rabinovich. Atlas: End-to-end 3d scene reconstruction from posed images. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VII 16*, pages 414–431. Springer, 2020.

[8] Julian Straub, Thomas Whelan, Lingni Ma, Yufan Chen, Erik Wijmans, Simon Green, Jakob J Engel, Raul Mur-Artal, Carl Ren, Shobhit Verma, et al. The replica dataset: A digital replica of indoor spaces. *arXiv preprint arXiv:1906.05797*, 2019.

[9] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler, faster, stronger. *arXiv preprint arXiv:2312.10035*, 2023.

[10] Saining Xie, Jiatao Gu, Demi Guo, Charles R Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16*, pages 574–591. Springer, 2020.

[11] Chandan Yeshwanth, Yueh-Cheng Liu, Matthias Nießner, and Angela Dai. Scannet++: A high-fidelity dataset of 3d indoor scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12–22, 2023.

[12] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15838–15847, 2021.

[13] Zhisheng Zhong, Jiequan Cui, Yibo Yang, Xiaoyang Wu, Xiaojuan Qi, Xiangyu Zhang, and Jiaya Jia. Understanding imbalanced semantic segmentation through neural collapse. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19550–19560, 2023.