

Supplementary Materials: GS²-GNeSF: Geometry-Semantics Synergy for Generalizable Neural Semantic Fields

Anonymous Authors

1 NETWORK ARCHITECTURE OF GCC

Global context compensation (GCC) utilizes a dual-branch encoder network to provide boundary compensation and semantic compensation. This encoder consists boundary branch and semantic branch. The boundary branch is responsible for encoding spatial details, which are low-level information, necessitating substantial channel capacity for encoding rich spatial details. Therefore, we utilize a wider channel but shallower convolutional network for the boundary branch, as shown in Table 1. Conversely, the semantic branch is tasked with outputting high-level semantic features that require a larger receptive field. Accordingly, we design a deeper convolutional network with fewer channels for the semantic branch, detailed in Table 1. Since the feature maps fed into the GCC have a low resolution (160×120), we don't reduce the resolution of the feature map within the encoder network. Each convolutional layer comprises a 2D convolution, a ReLU activation function, and instance normalization [5]. The second last layer in each branch outputs the feature maps serving as compensation features (boundary-aware or class-specific features) produced by that branch.

2 ADDITIONAL RESULTS AND ANALYSIS

2.1 Sampling Efficiency

To validate the efficiency of geometry-aware sampling, we conduct experiments to investigate the impact of the number of sampling points on model performance. As shown in Figure 1, even with only four sampling points, the model's semantic segmentation result is close to the best outcomes. Meanwhile, although the rendering quality decreases when using only four points, the PSNR remains above 29.5 dB, which is much better than S-Ray [3] (26.57 dB) and GNeSF [1] (24.44 dB), both of which used at least 128 sampling points. This result further confirms the effectiveness of our geometry-aware sampling strategy.

Moreover, increasing the number of sampling points to 16 and 32 does not enhance performance. This suggests that the model is not sensitive to the number of sampling points. Additionally, from Figure 1, we observe that RGB rendering is more sensitive to the number of sampling points compared to semantic prediction.

2.2 Qualitative Results with Finetuning

Following S-Ray [3], we evaluate our model's performance in finetuning setting. Specifically, we fine-tune our generalized model for a limited number of steps, 20k steps, on each unseen scene before evaluation. Figure 2 displays the segmentation results of our model compared to S-Ray[3] and GNeSF [1] under a finetuning setting. We observe that by finetuning with limited time (20k iterations for training), our model achieves better results compared to S-Ray [3] and GNeSF [1] especially in object's boundaries.

Description	Layer	Input Channels	Output Channels	K	Activation	Normalization
Common Input	-	32	-	-	-	-
Semantic Branch	Conv1	32	32	3×3	ReLU	Instance
	Conv2	32	32	3×3	ReLU	Instance
	Conv3	32	32	3×3	ReLU	Instance
	Conv4	32	32	3×3	ReLU	Instance
	Conv5	32	32	3×3	ReLU	Instance
	Conv6	32	32	3×3	ReLU	Instance
	Conv7	32	64	3×3	ReLU	Instance
	Conv8	64	64	3×3	ReLU	Instance
	Conv9	64	64	3×3	ReLU	Instance
	Conv10	64	64	3×3	ReLU	Instance
	Conv11	64	64	3×3	ReLU	Instance
	Conv12	64	64	3×3	-	-
Boundary Branch	Output	64	num_cls	1×1	-	-
	Conv1	32	128	3×3	ReLU	Instance
	Conv2	128	128	3×3	ReLU	Instance
	Conv3	128	128	3×3	ReLU	Instance
	Conv4	128	128	3×3	ReLU	Instance
	Conv5	128	128	3×3	ReLU	Instance
	Conv6	128	64	3×3	-	-
	Output	64	1	1×1	-	-

Table 1: Dual-Branch Neural Network Architecture. Here, num_cls denotes the number of predicted classes, K stands for kernel size.

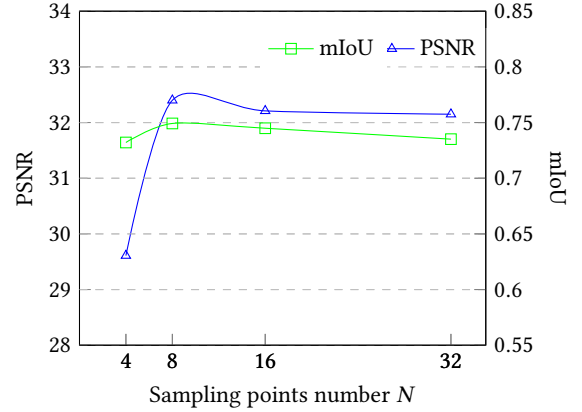


Figure 1: Sampling Efficiency on ScanNet [2]. We conduct experiments to explore the impact of the number of sampling points on model performance.

2.3 Additional Ablation Study on Robust Geometric Prior Generation

To investigate the effectiveness of using 3D features over 2D features in constructing the cost volume for the target view, we conduct an ablation study comparing our approach with methods like NeRF-SDP [6] and Garf [4]. These methods regress coarse depth maps on target views to guide sampling, similar to our approach. However, unlike their reliance on 2D feature maps from nearby reference views to build the cost volume for the target view, our robust geometric prior generation module utilizes 3D features extracted



Figure 2: Segmentation results of finetuning on unseen scenes.

from these reference views. As shown in Figure 3, cost volumes built with 3D features provides the model a more robust geometric prior, as evidenced by improved depth maps, superior quality RGB renderings, and enhanced semantic segmentation outcomes.

2.4 Analysis of Loss Weights

Each component of the loss function, including RGB loss, depth loss, semantic loss, and boundary loss as described in Eq. 9 of the main paper, contributes to the model’s training process. Therefore, we experiment with varying the weighting coefficients λ_{rgb} , λ_{depth} , λ_{sem} , and λ_{bon} on ScanNet dataset, to assess their individual and combined effects on model performance. We vary one coefficient at a time while keeping the others fixed at their default values and present the results in Table 2. The default values are $\lambda_{rgb} = 0.75$, $\lambda_{depth} = 0.8$, $\lambda_{sem} = 0.25$, and $\lambda_{bon} = 5$.

It is noteworthy that λ_{rgb} exhibits higher sensitivity than other loss weights. Reducing it to 0.25 significantly lowers both PSNR and mIoU compared to adjustments on other weights. This highlights RGB loss’s key role in maintaining image quality and semantic accuracy. Conversely, increasing λ_{rgb} to 1.00 slightly improves PSNR but does not significantly affect mIoU.

Moreover, variations in λ_{depth} show that reducing the depth weight to 0.40 or 0.60 degrades the performance in mIoU, emphasizing the importance of depth information in achieving higher scene understanding accuracy. Meanwhile, increasing λ_{depth} from 0.8 to 1.0 results in a slight improvement in rendering quality, however, the quality of segmentation experience a minor decline.

From the results presented in Table 2, increasing the semantic loss weight λ_{sem} from the default value of 0.25 to 0.75 and 1.00 results in mIoU changes from 74.97 to 74.05 and 74.55, respectively. The minimal impact and slight downward trend in semantic segmentation accuracy suggest that the model performance is quite robust to changes in the semantic loss.

Regarding λ_{bon} , when it increases from 1 to 5, both rendering and segmentation outcomes improve, demonstrating the effectiveness of the boundary detection introduced in the global context compensation module. However, when increased to 10, the model performance remains stable with minimal changes.

In the experiments, we determine the best combination of loss weights by conducting grid search on the ScanNet validation set with the goal of maximizing the semantic segmentation performance (mIoU). This process yields the optimal values: $\lambda_{rgb} = 0.75$,

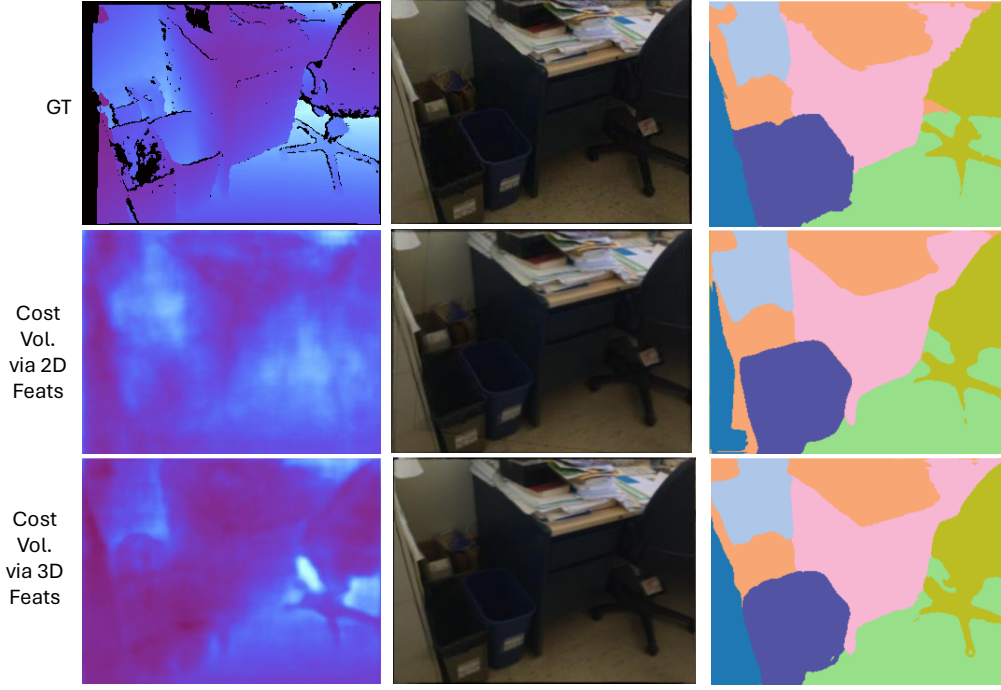


Figure 3: Quality comparison between different methods of constructing cost volumes.

λ_{rgb}	λ_{depth}	λ_{sem}	λ_{bon}	PSNR	mIoU
0.75	0.8	0.25	5	32.22	74.97
0.25	0.8	0.25	5	29.85	73.50
1.00	0.8	0.25	5	33.10	74.20
0.75	0.40	0.25	5	31.98	73.71
0.75	0.60	0.25	5	32.08	74.30
0.75	1.00	0.25	5	32.28	74.45
0.75	0.8	0.75	5	31.77	74.05
0.75	0.8	1.00	5	32.22	74.55
0.75	0.8	0.25	1	31.25	73.88
0.75	0.8	0.25	10	32.33	74.55

Table 2: Ablation Study on Loss Weightings for GS²-GNeSF on ScanNet [2].

$\lambda_{depth} = 0.8$, $\lambda_{sem} = 0.25$, and $\lambda_{bon} = 5$ on ScanNet dataset. We apply these optimal values to the Replica dataset as well.

REFERENCES

- [1] Hanlin Chen, Chen Li, Mengqi Guo, Zhiwen Yan, and Gim Hee Lee. 2024. GNeSF: Generalizable Neural Semantic Fields. *Advances in Neural Information Processing Systems* 36 (2024).
- [2] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. 2017. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5828–5839.
- [3] Fangfu Liu, Chubin Zhang, Yu Zheng, and Yueqi Duan. 2023. Semantic Ray: Learning a Generalizable Semantic Field with Cross-Reprojection Attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 17386–17396.
- [4] Yue Shi, Dingyi Rong, Bingbing Ni, Chang Chen, and Wenjun Zhang. 2022. Garf: Geometry-aware generalized neural radiance field. *arXiv preprint arXiv:2212.02280* (2022).

- [5] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2017. Improved texture networks: Maximizing quality and diversity in feed-forward stylization and texture synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6924–6932.
- [6] Qiuwen Wang, Shuai Guo, Haoning Wu, Rong Xie, Li Song, and Wenjun Zhang. 2023. NeRF-SDP: Efficient Generalizable Neural Radiance Field with Scene Depth Perception. In *Proceedings of the 5th ACM International Conference on Multimedia in Asia*. 1–7.