# GP-NeRF: Generalized Perception NeRF for Context-Aware 3D Scene Understanding

## Supplementary Material

### 7.1. Implementation of our Transformers

We provide a simple and efficient pytorch pseudo-code to implement the attention operations in the field-aggregation, ray-aggregation transformer blocks in Alg. 1, 2. We use ray features to generate attention maps $A_{field}$ and $A_{ray}$, and reuse them to construct semantic-embedding field as well as semantic features rendering.

**Algorithm 1:** Field-Aggregation Transformer

**Input:**
$\boldsymbol{X}_0 \rightarrow$ coordinate aligned features $(N_{\text{rays}}, N_{\text{pts}}, D_{rgb})$
$\boldsymbol{X}_{rgb} \rightarrow$ epipolar view Ray feats $(N_{\text{rays}}, N_{\text{pts}}, N_{\text{views}}, D_{rgb})$
$\boldsymbol{X}_{sem} \rightarrow$ epipolar view Sem feats $(N_{\text{rays}}, N_{\text{pts}}, N_{\text{views}}, D_{sem})$

$\boldsymbol{\Delta d} \rightarrow$ relative directions $(N_{\text{rays}}, N_{\text{pts}}, N_{\text{views}}, 3)$
**Network:** $f_Q, f_K, f_V, f_P, f_A, f_{rgb} \rightarrow$ MLP layers
**Output:** $\boldsymbol{S}^{3D}_{rgb}, \boldsymbol{S}^{3D}_{sem}$

**Forward:** Red for semantic-embedding field aggregation

$\boldsymbol{Q} = f_Q(\boldsymbol{X}_0), \boldsymbol{K} = f_K(\boldsymbol{X}_{rgb}), \boldsymbol{V} = f_V(\boldsymbol{X}_{rgb})$
$\boldsymbol{P}_{field} = f_P(\boldsymbol{\Delta d})$
$\boldsymbol{A}_{field} = \boldsymbol{K} - \boldsymbol{Q}[:, :\text{None}, :] + \boldsymbol{P}$
$\boldsymbol{A}_{field} = \text{softmax}(\boldsymbol{A}, \text{dim} = -2)$
$\boldsymbol{A}'_{field} = \boldsymbol{A}_{field} \cdot \text{repeat\_interleave}(4)$
$\boldsymbol{P}'_{field} = \boldsymbol{P} \cdot \text{repeat\_interleave}(4)$
$\boldsymbol{S}^{3D}_{rgb} = ((\boldsymbol{V} + \boldsymbol{P}) \cdot \boldsymbol{A}) \cdot \text{sum}(\text{dim} = 2)$
$\boldsymbol{S}^{3D}_{rgb} = f_{rgb}(\boldsymbol{S}^{3D}_{rgb})$
$\boldsymbol{S}^{3D}_{sem} = ((\boldsymbol{X}_{sem} + \boldsymbol{P}'_{field}) \cdot \boldsymbol{A}'_{field}) \cdot \text{sum}(\text{dim} = 2)$

**Algorithm 2:** Ray-Aggregation Transformer

**Input:**
$\boldsymbol{X}^{rgb}_0 \rightarrow$ coordinate aligned rgb features $(N_{\text{rays}}, N_{\text{pts}}, D_{rgb})$
$\boldsymbol{X}^{sem}_0 \rightarrow$ coordinate aligned sem features $(N_{\text{rays}}, N_{\text{pts}}, D_{sem})$
$\boldsymbol{x} \rightarrow$ point coordinates (after PE) $(N_{\text{rays}}, N_{\text{pts}}, D_{rgb})$
$d \rightarrow$ target view direction (after PE) $(N_{\text{rays}}, N_{\text{pts}}, D_{rgb})$
**Network:** $f_Q, f_K, f_V, f_P, f_A, f_{rgb}, f_{sem} \rightarrow$ MLP layers
**Output:** $\boldsymbol{S}^{2D}_{rgb}, \boldsymbol{S}^{2D}_{sem}$

**Forward:** Red for semantic-embedding field aggregation

$\boldsymbol{X}^{rgb}_0 = f_P(\text{concat}(\boldsymbol{X}^{rgb}_0, d, x))$
$\boldsymbol{Q} = f_Q\left(\boldsymbol{X}^{rgb}_0\right), \boldsymbol{K} = f_K\left(\boldsymbol{X}^{rgb}_0\right), \boldsymbol{V} = f_V\left(\boldsymbol{X}^{rgb}_0\right)$
$\boldsymbol{A}_{ray} = \text{matmul}\left(\boldsymbol{Q}, \boldsymbol{K}^T\right) / \sqrt{D}$
$\boldsymbol{A}_{ray} = \text{softmax}(\boldsymbol{A}_{ray}, \text{dim} = -1)$
$\boldsymbol{A}'_{ray} = \boldsymbol{A}_{ray} \cdot \text{repeat\_interleave}(4)$
$\boldsymbol{S}^{2D}_{rgb} = \text{matmul}(\boldsymbol{V}, \boldsymbol{A}_{ray})$
$\boldsymbol{S}^{2D}_{rgb} = f_{rgb}(\boldsymbol{S}^{2D}_{rgb})$
$\boldsymbol{S}^{2D}_{sem} = \text{matmul}(\boldsymbol{X}^{sem}_0, \boldsymbol{A}'_{ray})$

Table 5. Quantitative results of reconstruction task in Replica[33] during instance segmentation setting.

| Scene | DM-NeRF PSNR↑ | DM-NeRF SSIM↑ | DM-NeRF LPIPS↓ | Ours PSNR↑ | Ours SSIM↑ | Ours LPIPS↓ |
|---|---|---|---|---|---|---|
| Office_0 | 40.66 | 0.972 | 0.07 | 39.25 | 0.984 | 0.027 |
| Office_2 | 36.98 | 0.964 | 0.115 | 36.01 | 0.974 | 0.042 |
| Office_3 | 35.34 | 0.955 | 0.078 | 36.02 | 0.982 | 0.027 |
| Office_4 | 32.95 | 0.921 | 0.172 | 32.75 | 0.94 | 0.085 |
| Room_0 | 34.97 | 0.94 | 0.127 | 34.29 | 0.972 | 0.049 |
| Room_1 | 34.72 | 0.931 | 0.134 | 36.45 | 0.968 | 0.043 |
| Room_2 | 37.32 | 0.963 | 0.115 | 34.75 | 0.960 | 0.085 |
| Average | 36.13 | 0.949 | 0.116 | $35.64_{0.49\downarrow}$ | $0.969_{0.02\uparrow}$ | $0.051_{0.065\downarrow}$ |

### 7.2. Reconstruction results in instance setting

During the novel view instance segmentation task, we evaluate our reconstruction results and compare them with SOTA method DM-NeRF[38]. As shown in Table 5, our approach surpasses DM-NeRF in terms of SSIM and LPIPS metrics by 0.02% and 0.065%, respectively. It demonstrates that contextual information from semantic features can enhance the geometry reconstruction in our jointly optimized field and rendering framework.

### 7.3. Few-step Finetuning Comparison

Tab. 6 presents a comparison of different models, showcasing their mIoU and finetuning times on the ScanNet [9] dataset, along with the AP75 metric in Replica [33]. We observe that by finetuning with limited time, our model is able to achieve a better perception accuracy than a well-trained per-scene optimized method, such as 3.45% in mIoU with Semantic-NeRF [52] and 3.7% in AP75 with DM-NeRF [38]. Specifically, we observe that our method surpasses Semantic-Ray, requiring only half as many finetuning steps, and improves the mIoU by 0.74%, which further demonstrates that our semantic embedding field with more discrimination successfully improves the generalized ability.

We further evaluate the above experiments in instance segmentation setting, shown in the bottom column in Tab. 6. Not surprising, compared with SOTA method DM-NeRF[38], we achieve better performance with only 4k training steps, by 3.7% in AP75.

### 7.4. Additional Visualization Results

Fig. 9 shows the additional qualitative results of semantic prediction and reconstruction.

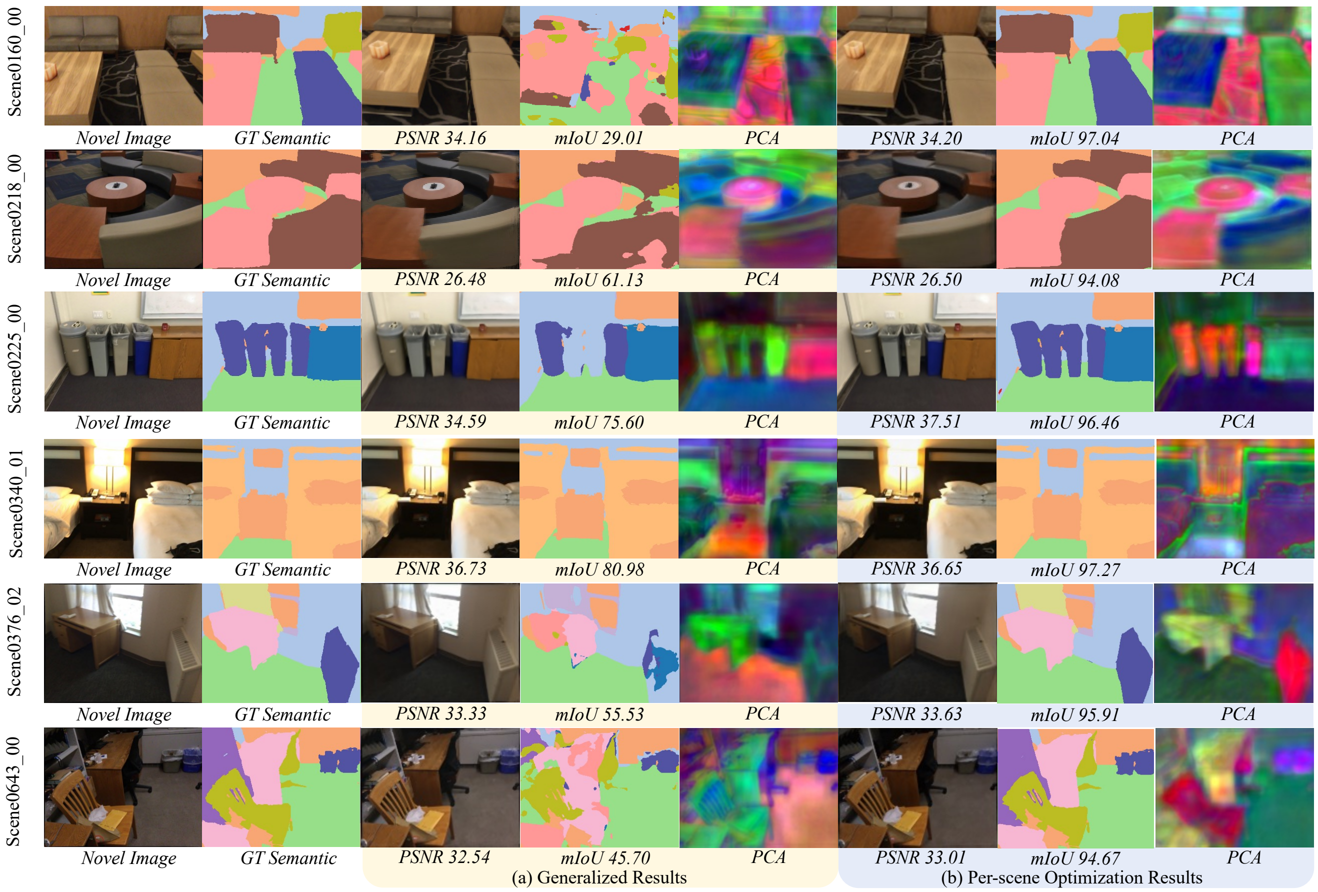


Figure 9. The visualization results in ScanNet[9]. Here we visualize the semantic as well as reconstruction results in both generalized and finetuning settings.

| Method | Train Step | Train Time | mIoU/AP75 |
|---|---|---|---|
| Semantic-NeRF [52] | 50k | ∼2h | 89.33 |
| MVSNeRF w/s-Ft | 5k | ∼20min | 52.02 |
| NeuRay [25] w/s-Ft | 5k | ∼32min | 79.23 |
| Semantic-Ray [20]-Ft | 5k | ∼20min | 92.04 |
| Ours-Ft | 2.5k | ∼20min | $92.78_{0.74\uparrow}$ |
| DM-NeRF | 200k | ∼2h | 81.03 |
| Ours-Ft | 4k | ∼30min | $84.73_{3.7\uparrow}$ |

Table 6. mIoU and training steps/time on ScanNet [9]. "w/ s" means adding a semantic head on the baseline architectures.