

Self-Supervised Implicit Attention Priors for Point Cloud Reconstruction

Supplementary Material

6. Further Experiments

6.1. Normal Estimation

Dataset and Metric: We adopt the evaluation protocol from [23], using the PCPNet dataset [11], which contains synthetic 3D shapes with a variety of surface characteristics, ranging from smooth regions to complex geometries with sharp features. Each shape is provided as a clean point cloud along with versions corrupted by Gaussian noise at three levels (0.12%, 0.6%, and 1.2% of the bounding box diagonal), as well as point clouds with non-uniform densities. Following [23], we report the oriented root mean square error (RMSE) of predicted normals (see Appendix 8.5 for details). Baseline methods and their corresponding results are adopted from [23] to ensure consistency and comparability.

Comparison We evaluate against a comprehensive set of baselines, including both classical and learning-based methods. Classical techniques include Principal Component Analysis (PCA) [14] and Locally Robust Regression (LRR) [47], each combined with three orientation propagation strategies: Minimum Spanning Tree (MST) [14], Sign Orientation Propagation (SNO) [38], and Orientation Determination Propagation (ODP) [32]. Learning-based baselines include AdaFit [50], HSurf-Net [22], PCPNet [11], SHS-Net [24], and NeuralGF [23].

Results. Table 5 reports the RMSE of oriented normal predictions across different noise levels and point density variations. Our method achieves the lowest error under the highest noise level (1.2%), indicating strong robustness to heavy corruption. It also performs competitively at moderate noise levels and under varying densities, ranking third overall in average RMSE behind NeuralGF and SHS-Net, the second of which is a fully supervised method. Notably, our approach outperforms several supervised baselines such as PCPNet and AdaFit, and consistently surpasses all classical methods by a significant margin. These results highlight our method’s ability to generalize well across challenging scenarios, despite not relying on supervised training signals.

6.2. Further Ablation Studies

To evaluate the impact of dictionary size and the cross-attention mechanism described in Section 3.1, we conduct a series of controlled ablation experiments.

We perform our analysis on the virus model from the self-similar dataset, where we expect local structure to

Table 5. RMSE of oriented normals on PCPNet dataset. Our method achieves competitive performance even when compared to supervised baselines.

METHOD	NOISE LEVEL				DENSITY		AVG
	None	0.12%	0.6%	1.2%	Stripe	Grad.	
PCA + MST	19.05	30.20	31.76	39.64	27.11	23.38	28.52
PCA + SNO	18.55	21.61	30.94	39.54	23.00	25.46	26.52
PCA + ODP	28.96	25.86	34.91	51.52	28.70	23.00	32.16
LRR + MST	43.48	47.58	38.58	44.08	48.45	46.77	44.82
LRR + SNO	44.87	43.45	33.46	45.40	46.96	37.73	41.98
LRR + ODP	28.65	25.83	36.11	53.89	26.41	23.72	32.44
AdaFit + MST	27.67	43.69	48.83	54.39	36.18	40.46	41.87
AdaFit + SNO	26.41	24.17	40.31	48.76	27.74	31.56	33.16
AdaFit + ODP	26.37	24.86	35.44	51.88	26.45	20.57	30.93
HSurf + MST	29.82	44.49	50.47	55.47	40.54	43.15	43.99
HSurf + SNO	30.34	32.34	44.08	51.71	33.46	40.49	38.74
HSurf + ODP	26.91	24.85	35.87	51.75	26.91	20.16	31.07
PCPNet	33.34	34.22	40.54	44.46	37.95	35.44	37.66
SHS-Net	10.28	13.23	25.40	35.51	16.40	17.92	19.79
NeuralGF	10.60	18.30	24.76	33.45	12.27	12.85	18.70
Ours	15.41	17.98	25.70	31.04	19.27	20.58	21.67

benefit from increased dictionary expressiveness. We vary the dictionary size across a range of values from 2 to 20 and measure reconstruction quality using the Chamfer Distance between the predicted distance field and the ground truth. Specifically, we sample the predicted implicit surface defined by the attentive signed distance function (SDF), convert it to a point cloud, and compute the distance to the ground-truth point cloud. Results are plotted in Fig. 10.

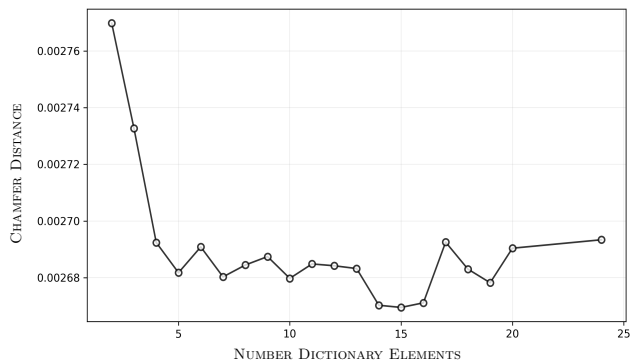


Figure 10. Plot shows the change in Chamfer Distance of the Neural-fields zero level-set against number of elements used within the dictionary.

We observe a clear trend: when the dictionary size is small (e.g., 2–4), the reconstruction is degraded. This is because the small dictionary primarily encodes coarse, global patterns, limiting expressiveness. As the dictionary size increases, reconstructions become progressively sharper and more faithful to the ground truth, indicating improved local pattern representation through richer token diversity. However, beyond a certain size, performance begins to saturate - this is accompanied by increasing similarity between tokens in the dictionary, suggesting redundancy. Based on this trade-off, we select a dictionary size of 16 for all main experiments, balancing accuracy and efficiency.

7. Experimental Details

7.1. Implementation and Environment

All experiments were conducted using PyTorch with PyTorch3D for geometric operations. Training is performed on a single NVIDIA RTX A5000 GPU, with each shape taking approximately 8–12 minutes to converge over 20,000 epochs.

7.2. Network Architecture

We adopt an MLP architecture similar to that used in NeuralGF. The neural field f_θ is modeled using an 8-layer MLP with hidden dimension 256 and a skip connection at the midpoint layer. We apply geometric initialization as described in [1], stabilizing the signed distance function near the zero level set.

To encode query coordinates, we use a sinusoidal positional encoder with 6 frequency bands. The encoded query is passed through a cross-attention module that interacts with a shared latent dictionary of geometric tokens. We use 8 attention heads in our multi-headed attention setup.

7.3. Cross-Attention Prior

The latent self-prior is implemented as a learnable embedding dictionary containing 16 tokens, initialized using QR decomposition of random matrices and updated via back-propagation. Cross-attention is applied between the encoded queries and dictionary tokens using multi-head attention, dynamically aggregating non-local geometric information across the shape.

7.4. Training Procedure

The training loss combines several self-supervised geometric terms, as detailed in the main paper; we use the following hyperparameters across all our experiments: $\alpha = 0.3$, $\beta = 10$, $\gamma = 1$, and $\delta = 0.01$. Training samples include both on-surface points from the input point cloud and off-surface points obtained via Gaussian

perturbation. We follow the procedure introduced in [23]; to generate training samples, we first normalize the input mesh and downsample to a maximum of 300,000 points. For each point, we compute the distance to its 50th nearest neighbor and use this value as a local scale parameter. We then generate noisy query points by applying Gaussian perturbation scaled by the local distance and a global factor (`dis_scale` = 0.15). For each query point, we identify its 64 nearest neighbors to construct local patches for geometric supervision. We generate up to 10 rounds of query points per shape, yielding a large, dense set of perturbed inputs and associated neighborhoods.

We train each shape independently using the Adam optimizer. We use a two-stage learning rate schedule: an initial linear warm-up phase followed by cosine annealing. During the first 10,000 iterations, the learning rate increases linearly from zero to the base learning rate of 1×10^{-4} . After the warm-up, the learning rate follows a cosine decay schedule until the end of training at 20,000 iterations. This approach encourages stable early training and smooth convergence.

8. Mesh Reconstruction Quality Metrics

To quantitatively evaluate the quality of the reconstructed 3D meshes (M_{REC}) against their corresponding ground truth meshes (M_{GT}), we employ a suite of established geometric metrics. For metrics requiring point cloud representations, we uniformly sample N_s points from the surfaces of both M_{GT} and M_{REC} . Unless otherwise specified, $N_s = 100,000$ for Chamfer and Hausdorff distances, and $N_s = 10,000$ for F-Score computation.

8.1. Chamfer Distance (CD)

The Chamfer Distance measures the average squared distance between closest point pairs across two point sets. Let $S_{\text{GT}} = \{\mathbf{p}_1, \dots, \mathbf{p}_{N_s}\}$ be the set of points sampled from M_{GT} , and $S_{\text{REC}} = \{\mathbf{q}_1, \dots, \mathbf{q}_{N_s}\}$ be the set of points sampled from M_{REC} . The Chamfer Distance is defined as:

$$d_{\text{CD}}(S_{\text{GT}}, S_{\text{REC}}) = \frac{1}{|S_{\text{GT}}|} \sum_{\mathbf{p} \in S_{\text{GT}}} \min_{\mathbf{q} \in S_{\text{REC}}} \|\mathbf{p} - \mathbf{q}\|_2^2 + \frac{1}{|S_{\text{REC}}|} \sum_{\mathbf{q} \in S_{\text{REC}}} \min_{\mathbf{p} \in S_{\text{GT}}} \|\mathbf{q} - \mathbf{p}\|_2^2 \quad (4)$$

where $\|\cdot\|_2$ denotes the Euclidean L2-norm. A lower CD value indicates a better alignment between the two point sets, signifying higher reconstruction accuracy in terms of average surface proximity.

8.2. Hausdorff Distance (HD)

The Hausdorff Distance captures the maximum discrepancy between two point sets. It is a more stringent metric than

CD as it is sensitive to outliers or localized large errors. Using the same point sets S_{GT} and S_{REC} as defined for CD, the Hausdorff Distance is given by:

$$d_{HD}(S_{GT}, S_{REC}) = \max \left\{ \sup_{p \in S_{GT}} \inf_{q \in S_{REC}} \|p - q\|, \sup_{q \in S_{REC}} \inf_{p \in S_{GT}} \|q - p\| \right\} \quad (5)$$

where \sup denotes the supremum (least upper bound) and \inf denotes the infimum (greatest lower bound). A lower HD value signifies a smaller maximum error between the surfaces.

8.3. F-Score (F_1)

The F-Score evaluates surface reconstruction quality by considering both precision and recall with respect to a distance threshold τ . Points P_{GT} are sampled from M_{GT} and P_{REC} from M_{REC} (with $N_s = 10,000$ samples for this metric). Precision (P) is the fraction of points in P_{REC} that are within distance τ of any point in P_{GT} :

$$P(\tau) = \frac{1}{|P_{REC}|} \sum_{q \in P_{REC}} \mathbb{I} \left(\min_{p \in P_{GT}} \|q - p\|_2 < \tau \right) \quad (6)$$

Recall (R) is the fraction of points in P_{GT} that are within distance τ of any point in P_{REC} :

$$R(\tau) = \frac{1}{|P_{GT}|} \sum_{p \in P_{GT}} \mathbb{I} \left(\min_{q \in P_{REC}} \|p - q\|_2 < \tau \right) \quad (7)$$

where $\mathbb{I}(\cdot)$ is the indicator function, returning 1 if the condition is true, and 0 otherwise. The F-Score is the harmonic mean of precision and recall:

$$F_1(\tau) = 2 \cdot \frac{P(\tau) \cdot R(\tau)}{P(\tau) + R(\tau)} \quad (8)$$

A higher F-Score (closer to 1) indicates better overall agreement between the surfaces, considering both completeness (recall) and correctness (precision).

8.4. Normal Consistency (NC)

Normal Consistency measures the alignment of surface normals between the reconstructed mesh M_{REC} and the ground truth mesh M_{GT} . This metric is crucial for assessing the smoothness and geometric detail preservation of the reconstructed surface. Let F_{REC} be the set of faces in M_{REC} . For each face $f_i \in F_{REC}$, let c_i be its centroid and \hat{n}_i be its unit normal vector. We find the corresponding face $f_j^* \in F_{GT}$ (the set of faces in M_{GT}) whose centroid c_j^* is closest to c_i :

$$c_j^* = \arg \min_{c_k \in C_{GT}} \|c_i - c_k\|_2 \quad (9)$$

where C_{GT} is the set of all face centroids in M_{GT} . Let \hat{n}_j^* be the unit normal of this closest ground truth face f_j^* . The Normal Consistency is then computed as the average of the absolute dot products of these corresponding normal pairs:

$$NC = \frac{1}{|F_{REC}|} \sum_{f_i \in F_{REC}} |\hat{n}_i \cdot \hat{n}_j^*| \quad (10)$$

The NC score ranges from 0 to 1, where 1 indicates perfect alignment of normals between the reconstructed mesh and the corresponding parts of the ground truth mesh. A higher NC score suggests that the reconstructed surface accurately captures the local orientation of the ground truth surface.

8.5. Normal Estimation Metric

The Oriented Root Mean Squared Error (RMSE_O) quantifies the angular deviation between estimated surface normals and ground truth normals, taking orientation into account. This metric is crucial in applications where the direction of normals affects downstream tasks such as rendering or shading. Let \hat{n}_i and n_i denote the unit ground-truth and predicted normals, respectively, for each of the I evaluation points. RMSE_O is computed as:

$$RMSE_O = \sqrt{\frac{1}{I} \sum_{i=1}^I (\arccos(\hat{n}_i \cdot n_i))^2} \quad (11)$$

The angular error is measured in degrees, ranging from 0° (perfect alignment) to 180° (opposite orientation). A lower RMSE_O indicates more accurate normal orientation estimation, highlighting the fidelity of the reconstruction process.

9. Qualitative Results Point2Mesh



Figure 11. We compare our approach with Point2Mesh [13] using the publicly available objects released by the Point2Mesh authors. We note that in general our approach produces surfaces which are smoother while retaining sharp features.

10. Qualitative Results on SRB

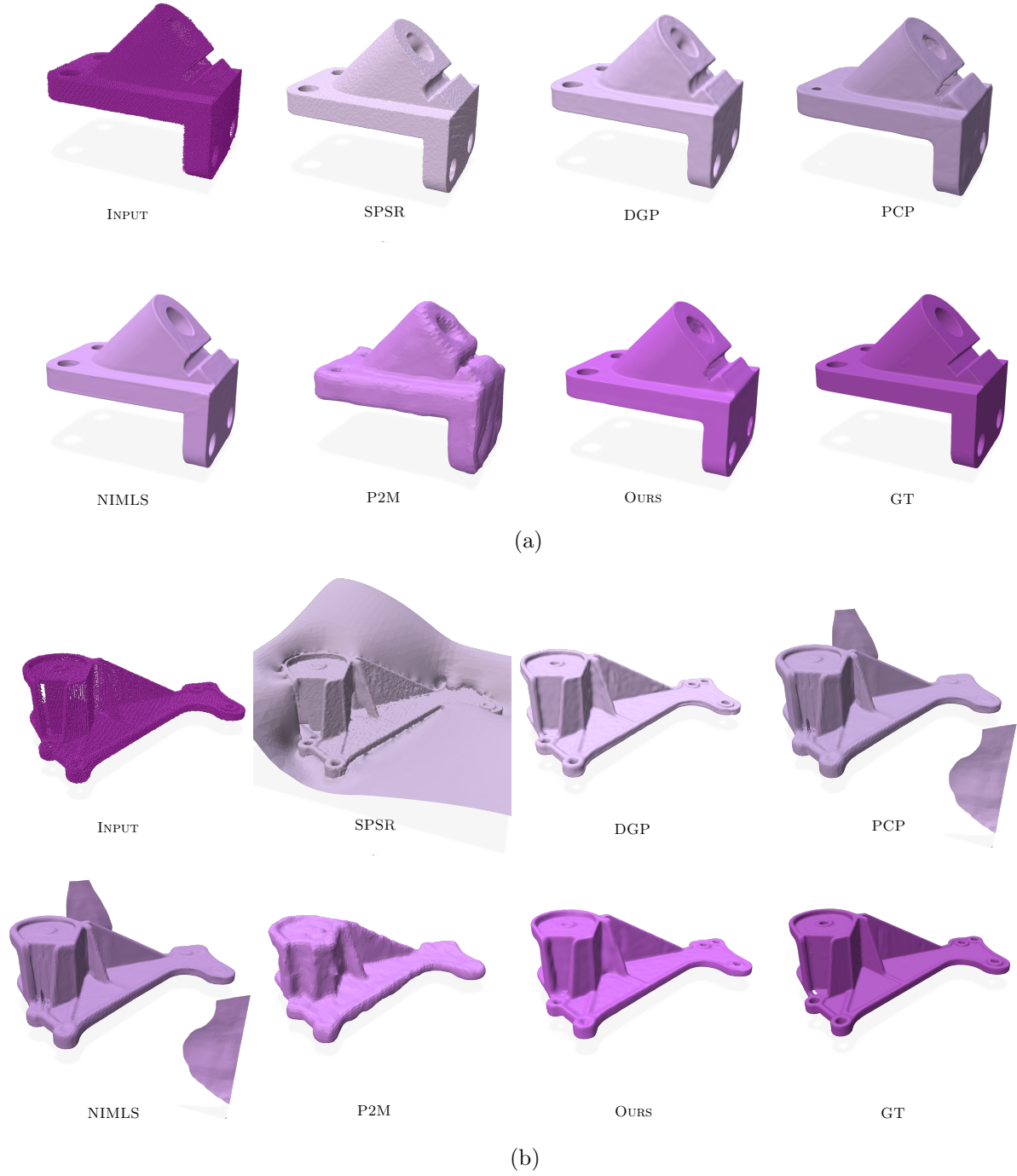
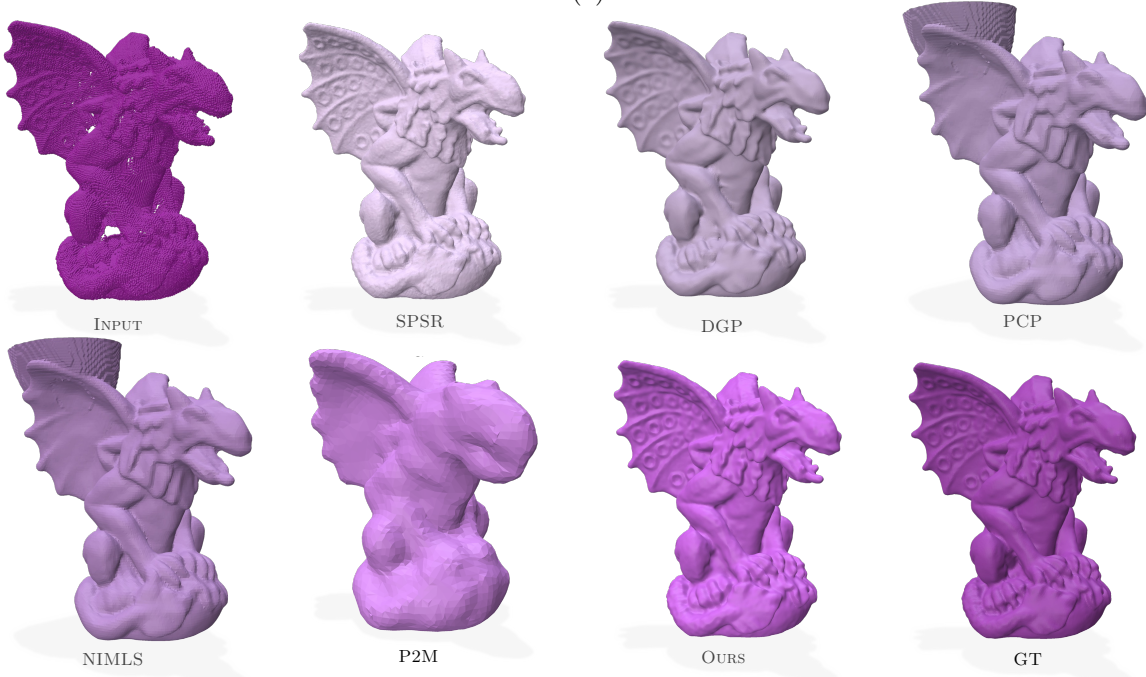


Figure 12. Shows the qualitative results of our method on objects from the *surface reconstruction benchmark* (SRB), compared against other reconstruction techniques. Methods are defined in Section 4.1.



(a)



(b)

Figure 13. Shows the qualitative results of our method on objects from the *surface reconstruction benchmark* (SRB), compared against other reconstruction techniques. Methods are defined in Section 4.1.

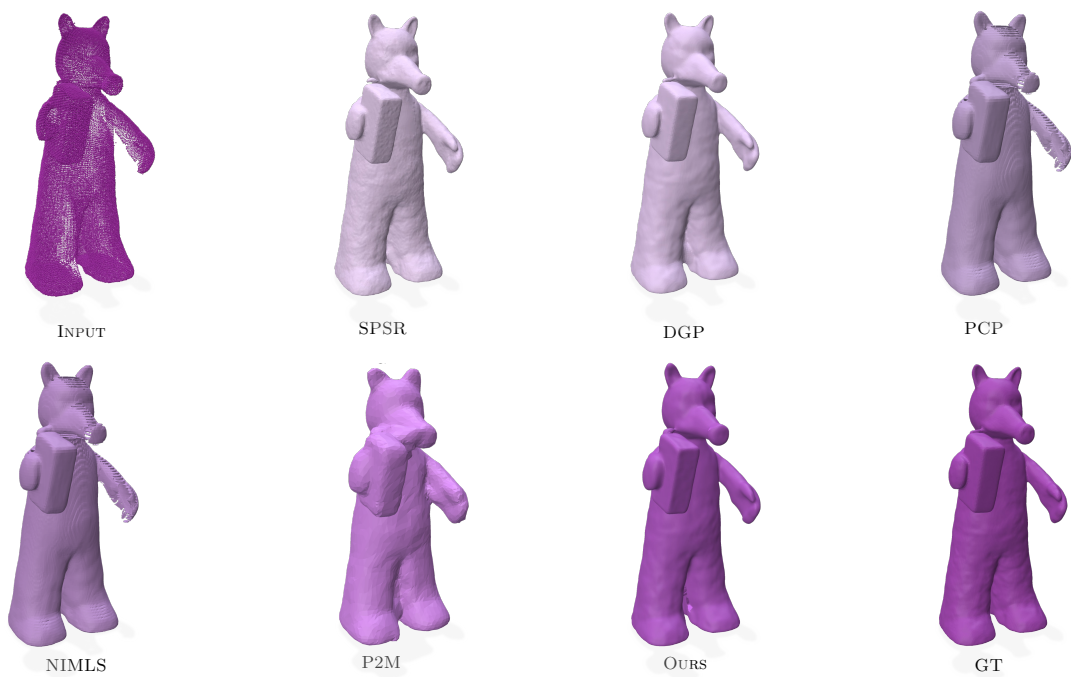


Figure 14. Shows the qualitative results of our method on objects from the *surface reconstruction benchmark* (SRB), compared against other reconstruction techniques. Methods are defined in Section 4.1.