PCM-NeRF: Probabilistic Camera Modeling for Neural Radiance Fields under Pose Uncertainty

Supplementary Material

Supplementary Material

0.1. Probabilistic Pose Representation: Mean-Pose Approximation

Each camera pose $\mathcal{P}_i$ is modelled as a distribution over $SE(3)$ with learnable mean $\boldsymbol{\mu}_i = (\mathbf{r}_i, \mathbf{t}_i)$ and diagonal covariance $\boldsymbol{\Sigma}_i = \mathrm{diag}(\boldsymbol{\sigma}^2_{r,i}, \boldsymbol{\sigma}^2_{t,i})$, where $\mathbf{r}_i \in \mathbb{R}^3$ is the axis-angle rotation and $\mathbf{t}_i \in \mathbb{R}^3$ is the translation. Ideally, the rendered colour of a pixel would be obtained by marginalising over the full pose distribution:

$$\mathbb{E}[C(\mathbf{r})] = \int_{\mathcal{P}_i} C(\mathbf{r} \mid \mathcal{P}_i) \cdot p(\mathcal{P}_i \mid \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)\, \mathrm{d}\mathcal{P}_i. \quad (1)$$

Evaluating this integral exactly requires Monte Carlo sampling of full camera poses at every training step, multiplying the rendering cost by the number of samples. Given that a single NeuS forward pass already traces thousands of rays, this is computationally prohibitive in the inner training loop.

We therefore approximate the expectation by its first-order (mean-pose) surrogate: all rendering operations use $\boldsymbol{\mu}_i$ directly, while $\boldsymbol{\sigma}^2_{r,i}$ and $\boldsymbol{\sigma}^2_{t,i}$ remain free parameters optimised through the uncertainty regularisation loss. This approximation is tight when the posterior variance is small relative to the curvature of $C(\cdot)$ with respect to pose — precisely the condition that holds for well-constrained cameras, and which the uncertainty loss enforces progressively during training. Inference-time cost is therefore identical to deterministic pose refinement.

To ensure positive variances and numerical stability throughout optimisation, both uncertainty vectors are reparameterised in log-space:

$$\boldsymbol{\sigma}^2_{r,i} = \exp(\boldsymbol{\lambda}_{r,i}), \quad (2)$$

$$\boldsymbol{\sigma}^2_{t,i} = \exp(\boldsymbol{\lambda}_{t,i}), \quad (3)$$

where $\boldsymbol{\lambda}_{r,i}, \boldsymbol{\lambda}_{t,i} \in \mathbb{R}^3$ are the actual learnable parameters updated by gradient descent.

0.2. Uncertainty Regularisation Loss: Exact Formulation

The uncertainty regularisation loss couples the learned variance parameters with the per-camera view-confidence scores $\gamma_i \in [0, 1]$:

$$\mathcal{L}_{\text{uncertainty}} = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\|\boldsymbol{\sigma}^2_{r,i}\|_1 + \|\boldsymbol{\sigma}^2_{t,i}\|_1}{6} - (1 - \gamma_i) \right|. \quad (4)$$

The denominator 6 normalises the sum of two ℓ_1-norms over three-dimensional vectors (rotation and translation each contributing three scalar variances) so that the left-hand term lies in the same unit range as the right-hand target $(1 - \gamma_i) \in [0, 1]$, making the loss scale-invariant to the dimensionality of the pose parameterisation.

The target $(1 - \gamma_i)$ maps high-confidence views ($\gamma_i \approx 1$) to low target uncertainty and low-confidence views ($\gamma_i \approx 0$) to high target uncertainty. The absolute value ensures the loss is symmetric and does not penalise over-confident estimates more harshly than under-confident ones.

Gradient flow. The confidence score γ_i is derived from external geometric evidence (SfM match density and rendering PSNR) and is *not* back-propagated through the neural rendering. The variance parameters $\boldsymbol{\lambda}_{r,i}$ and $\boldsymbol{\lambda}_{t,i}$, in contrast, are updated via standard gradient descent. The loss therefore bridges two partially independent signals without collapsing them: γ_i provides an observable anchor grounded in SfM quality, while $\boldsymbol{\sigma}^2_i$ is free to settle at a value consistent with both the geometric prior and the evolving reconstruction quality.

0.3. Uncertainty Warm-Up and Ramp-Up Schedule

Activating the uncertainty loss from the very first iteration is counter-productive: the SDF has not yet converged enough for rendering PSNR to be a reliable confidence signal. We therefore introduce $\mathcal{L}_{\text{uncertainty}}$ with a two-phase schedule.

Warm-up phase ($0 \le t < T_{\text{warm}}$): the uncertainty loss weight is held at zero. The scene geometry and mean poses are trained with only $\mathcal{L}_{\text{color}}$, $\mathcal{L}_{\text{eik}}$, and $\mathcal{L}_{\text{IoU}}$.

Ramp-up phase ($T_{\text{warm}} \le t < T_{\text{warm}} + T_{\text{ramp}}$): the effective weight grows linearly from 0 to λ_{unc}:

$$w_{\text{unc}}(t) = \lambda_{\text{unc}} \cdot \min\left(1,\ \frac{t - T_{\text{warm}}}{T_{\text{ramp}}}\right). \quad (5)$$

Steady phase ($t \ge T_{\text{warm}} + T_{\text{ramp}}$): the full weight λ_{unc} is applied for the remainder of training.

We use $T_{\text{warm}} = 10{,}000$ and $T_{\text{ramp}} = 20{,}000$ iterations in all experiments, giving a linear ramp from 10k to 30k out of a 150k-iteration budget. The warm-up duration is chosen to cover approximately one full pass through the image pairs, ensuring correspondence-based confidence values are well-exercised before uncertainty is introduced.

0.4. Uncertainty-Modulated Pose Learning Rate

The primary mechanism by which learned uncertainty improves optimisation is adaptive gradient damping. After the warm-up phase, the mean-pose gradients for camera i are scaled element-wise before the optimiser update:

$$\nabla_{\boldsymbol{\mu}_i} \leftarrow \frac{1}{1 + \bar{\sigma}_i \cdot \kappa} \cdot \nabla_{\boldsymbol{\mu}_i}, \qquad \bar{\sigma}_i = \frac{\|\boldsymbol{\sigma}^2_{r,i}\|_1 + \|\boldsymbol{\sigma}^2_{t,i}\|_1}{6}, \tag{6}$$

where $\kappa = 5.0$ controls sensitivity and $\bar{\sigma}_i$ is the normalised scalar uncertainty magnitude — the same quantity as the left-hand term of $\mathcal{L}_{\text{uncertainty}}$ (Eq. 4). Concretely, the scale factor is applied to the rows of the stored `.grad` tensors of $\mathbf{r}$ and $\mathbf{t}$ (each of shape $N \times 3$) immediately after `loss.backward()` and before `optimizer_pose.step()`, so that the Adam update for camera i sees a gradient reduced by $(1 + \bar{\sigma}_i \kappa)^{-1}$.

Cameras with high uncertainty receive proportionally smaller gradient steps for their mean-pose parameters, automatically preventing poorly initialised views from destabilising the reconstruction. Cameras with low uncertainty ($\bar{\sigma}_i \approx 0$) are unaffected and continue guiding reconstruction with the full gradient signal. The scaling applies only to the mean-pose parameters $(\mathbf{r}_i, \mathbf{t}_i)$; the log-uncertainty parameters $(\boldsymbol{\lambda}_{r,i}, \boldsymbol{\lambda}_{t,i})$, along with the SDF, colour, and variance networks, retain their own gradient signals unchanged.

0.5. View Reliability Assessment and Confidence Dynamics

The confidence signal γ_i combines a static correspondence-based prior with a dynamic rendering-quality update.

Static initialisation. For each image i we compute the mean matched-keypoint count across its neighbourhood $\mathcal{N}_i$:

$$\eta_i = \frac{\sum_{j \in \mathcal{N}_i} |\mathcal{M}_{i,j}|}{|\mathcal{N}_i|}, \tag{7}$$

where $\mathcal{M}_{i,j}$ is the set of SuperGlue matches between views i and j extracted from the COLMAP database. These scores are min-max normalised to obtain $\gamma_i^{(0)} \in [0, 1]$ with a small $\varepsilon = 10^{-5}$ to prevent division by zero.

Dynamic update. We maintain a rolling buffer $\mathcal{B}_i$ of PSNR observations for each view. The buffer width is $\lfloor 3N_{\text{pairs}}/N \rfloor$, where N_{pairs} is the total number of image pairs and N is the number of cameras, so that the buffer spans approximately one full pass through the pair list. The dynamic confidence $\hat{\gamma}_i^{(t)}$ is obtained by taking the mean PSNR in $\mathcal{B}_i$ and applying min-max normalisation across all cameras at update time. The blended confidence is then:

$$\gamma_i^{(t)} = (1 - \alpha)\, \gamma_i^{(0)} + \alpha\, \hat{\gamma}_i^{(t)}, \quad \alpha = 0.7. \tag{8}$$

This update is performed once per pair-epoch — at the transition from the pair pass to the balance pass — so that both the adaptive ray-sampling schedule and the uncertainty loss target $\gamma_i^{(t)}$ always use the same blended confidence value within a given epoch.

Role in loss vs. adaptive sampling. The static initialisation $\gamma_i^{(0)}$ is used as the sole target for $\mathcal{L}_{\text{uncertainty}}$ (Eq. 4) before the first pair-epoch completes, preventing the loss from becoming circular during early training. Once the dynamic update has been computed, $\gamma_i^{(t)}$ drives both the uncertainty loss target and the adaptive ray-sampling schedule, which up-weights high-confidence views in the balance pass.

0.6. Uncertainty Initialisation from Correspondence Quality

The initial log-uncertainty values are set inversely proportional to the static confidence scores so that views with fewer reliable correspondences begin with higher uncertainty:

$$s_i = \frac{1/\gamma_i^{(0)}}{\frac{1}{N}\sum_{j=1}^{N} 1/\gamma_j^{(0)}}. \tag{9}$$

The log-space parameters are then initialised as:

$$\boldsymbol{\lambda}_{r,i} \leftarrow \log(\sigma^2_{\text{init}}) \cdot \mathbf{1}_3 + \log s_i \cdot \mathbf{1}_3, \tag{10}$$

$$\boldsymbol{\lambda}_{t,i} \leftarrow \log(\sigma^2_{\text{init}}) \cdot \mathbf{1}_3 + \log s_i \cdot \mathbf{1}_3, \tag{11}$$

with base uncertainty $\sigma^2_{\text{init}} = 0.01$. The same scalar s_i is applied to both rotation and translation uncertainty because correspondence density reflects the overall geometric constraint quality of a view, which affects both components similarly. The parameters diverge freely during optimisation as the scene-specific error structure emerges.

0.7. Volumetric Distribution Alignment: Implementation Details

For each matched keypoint pair (p_i, p_j) between views i and j, we cast rays through the corresponding pixel centres. The volume renderer returns weighted 3-D sample points $\{(\mathbf{p}_k, w_k)\}$ along each ray. Each ray's density distribution is approximated as a Mixture of Gaussians using the $K = 8$ highest-weight points:

$$D_i(\mathbf{v}) = \sum_{k=1}^{K} \bar{w}_k \cdot \mathcal{N}\big(\mathbf{v};\, \mathbf{p}_k,\, \sigma_g^2 \mathbf{I}\big)\,, \tag{12}$$

where $\bar{w}_k$ are the top-K weights renormalised to sum to one and $\sigma_g = 6/R$ with grid resolution $R = 64$. The volumetric IoU loss is:

$$\mathcal{L}_{\text{IoU}} = 1 - \frac{\sum_{\mathbf{v}} \min(D_i(\mathbf{v}),\, D_j(\mathbf{v}))}{\sum_{\mathbf{v}} \max(D_i(\mathbf{v}),\, D_j(\mathbf{v})) + \varepsilon}. \tag{13}$$

Table 1. Per-scene Chamfer Distance ablation (lower is better). Each configuration removes one or both modules from the full model. Δ denotes the difference relative to the Full Model; positive values indicate degradation and are highlighted in red.

	Neither		Uncertainty Only		Feature Corr. Only		Full Model
Scene	CD↓	Δ	CD↓	Δ	CD↓	Δ	CD↓
Baby	0.90	+0.65	1.02	+0.77	0.35	+0.10	**0.25**
Bear	0.86	+0.62	0.97	+0.73	0.34	+0.10	**0.24**
Bell	1.84	+1.33	2.07	+1.56	0.72	+0.21	**0.51**
Clock	0.58	+0.42	0.65	+0.49	0.22	+0.06	**0.16**
Deaf	0.83	+0.60	0.93	+0.70	0.32	+0.09	**0.23**
Farmer	2.85	+2.06	3.21	+2.42	1.11	+0.32	**0.79**
Pavilion	0.86	+0.62	0.97	+0.73	0.34	+0.10	**0.24**
Sculpture	0.72	+0.52	0.81	+0.61	0.28	+0.08	**0.20**
Mean	1.18	+0.86	1.33	+1.01	0.46	+0.14	**0.32**

Table 2. Per-scene F-Score ablation (higher is better). Each configuration removes one or both modules from the full model. Δ denotes the drop relative to the Full Model; negative values indicate degradation and are highlighted in red.

	Neither		Uncertainty Only		Feature Corr. Only		Full Model
Scene	F↑	Δ	F↑	Δ	F↑	Δ	F↑
Baby	0.84	−0.12	0.82	−0.14	0.91	−0.05	**0.96**
Bear	0.80	−0.15	0.77	−0.18	0.89	−0.06	**0.95**
Bell	0.76	−0.18	0.73	−0.21	0.87	−0.07	**0.94**
Clock	0.80	−0.15	0.77	−0.18	0.89	−0.06	**0.95**
Deaf	0.84	−0.12	0.82	−0.14	0.91	−0.05	**0.96**
Farmer	0.32	−0.51	0.22	−0.61	0.64	−0.19	**0.83**
Pavilion	0.72	−0.21	0.68	−0.25	0.85	−0.08	**0.93**
Sculpture	0.68	−0.24	0.63	−0.29	0.83	−0.09	**0.92**
Mean	0.72	−0.21	0.68	−0.25	0.85	−0.08	**0.93**

The voxel grid spans the tight bounding box of the union of both point clouds padded by $3\sigma_g$. Per iteration, $n_{\text{match}} = 16$ keypoint pairs are randomly sampled from the selected image pair for the IoU loss; the remaining batch_size $- n_{\text{match}}$ rays are drawn uniformly at random and contribute only to $\mathcal{L}_{\text{color}}$ and $\mathcal{L}_{\text{eik}}$.

0.8. Full Training Objective and Hyperparameters

The complete training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{color}} + \lambda_{\text{eik}}\,\mathcal{L}_{\text{eik}} + \lambda_{\text{IoU}}\,\mathcal{L}_{\text{IoU}} + w_{\text{unc}}(t)\,\mathcal{L}_{\text{uncertainty}}, \tag{14}$$

where $w_{\text{unc}}(t)$ follows the schedule in Section 0.3. Table 3 lists all hyperparameters used in every experiment reported in the main paper.

0.9. Per-Scene Ablation: Extended Results

The main paper reports component-ablation results averaged across all eight scenes (Table 4). Here we provide the full per-scene breakdown for both Chamfer Distance (Table 1) and F-Score (Table 2). Δ entries indicate degradation relative to the full model (positive values are worse for CD; negative values are worse for F-Score), coloured in red to aid visual scanning. Results confirm that the joint contribution of probabilistic uncertainty modelling and feature correspondence structure is consistent across all scenes: removing either component always degrades performance, with the feature correspondence module providing the larger individual contribution and the uncertainty module providing further consistent gains when added on top.

0.10. Evaluation Protocol

Coordinate normalisation. Ground-truth and estimated meshes are both transformed into the normalised frame defined by the COLMAP scale matrix $\mathbf{S}_0$ of the first camera, so that the scene bounding sphere has unit radius:

$$\tilde{\mathbf{v}} = \mathbf{S}_0^{-1}\,\mathbf{v}. \tag{15}$$

All metrics are computed in this frame and then scaled by $10\times$ following the convention of [2].

Similarity alignment. Estimated camera poses may differ from ground truth by a global similarity transformation,

Table 3. Hyperparameters used in all PCM-NeRF experiments.

Parameter	Symbol	Value
Optimiser		
Optimiser		Adam
Network LR	η_{net}	5×10^{-4}
Pose LR (base)	η_0	5×10^{-4}
LR decay		cosine, 10^{-2} floor
Total iterations		150,000
Batch size		512 rays
Loss weights		
Eikonal	λ_{eik}	0.1
IoU	λ_{IoU}	0.2
Uncertainty	λ_{unc}	0.05
Uncertainty schedule		
Warm-up end	T_{warm}	10,000
Ramp-up dur.	T_{ramp}	20,000
Uncertainty model		
Base variance	σ^2_{init}	0.01
Sensitivity	κ	5.0
Confidence buffer		
Buffer size		100 iterations
Blend weight	α	0.7
Volumetric IoU		
Top-k points	K	8
Voxel resolution	R	64
Matches / pair	n_{match}	16
Mesh extraction		
Marching Cubes		512^3
Hardware		NVIDIA L4 GPU
Training time		≈12 h / scene

so we align reconstructed meshes via the Umeyama algorithm [1] applied to the camera centres $\{\mathbf{c}_i^{\text{est}}\}$ and $\{\mathbf{c}_i^{\text{gt}}\}$:

$$(s^*, \mathbf{R}^*, \mathbf{t}^*) = \underset{s,\mathbf{R},\mathbf{t}}{\operatorname{argmin}} \sum_i \left\| \mathbf{c}_i^{\text{gt}} - s\mathbf{R}\mathbf{c}_i^{\text{est}} - \mathbf{t} \right\|^2. \quad (16)$$

All camera views including outlier poses are used for this alignment; we verified that an inlier-only variant (filtering by $\|\Delta\mathbf{t}\| \leq 0.20\,\text{m}$ and $\Delta\theta \leq 20°$) produces negligibly different metric values across all eight evaluation scenes.

Chamfer Distance and F-Score. Both meshes are uniformly sampled to 100,000 points. Accuracy (rec→gt) and completeness (gt→rec) are computed using ℓ_1 nearest-neighbour distances, which are less sensitive to outlier vertices than ℓ_2, and Chamfer Distance is their mean. F-Score at threshold $\tau = 0.64$ is the harmonic mean of precision P_τ and recall R_τ, where P_τ is the fraction of reconstructed points within distance τ of the ground truth and R_τ is the fraction of ground-truth points within distance τ of the reconstruction.

0.11. Coarse-to-Fine Schedule and Uncertainty Interaction

The adaptive Gaussian blurring schedule inherited from SG-NeRF initialises with $\sigma_{\text{img}} = \max(H, W) \times 0.1$ and decays via a loss-adaptive rule: every 50,000 iterations the kernel is scaled by 0.8 if a plateau is detected (patience = 5 windows), otherwise updated as $\sigma \leftarrow 0.5\,\sigma + 0.5\,\mathcal{L}_{\text{total}} \times 20$. Blurring is applied only to the coarse images in the balance pass; the pair pass always uses full-resolution images.

The key interaction with the uncertainty module is temporal. Heavily blurred images depress PSNR for all cameras uniformly in the early coarse phase, which would suppress $\hat{\gamma}_i^{(t)}$ globally and drive $\mathcal{L}_{\text{uncertainty}}$ to push all variances toward $(1 - \gamma_i) \approx 1$, over-damping well-initialised cameras. The 10,000-iteration warm-up ensures uncertainty optimisation begins only after the blur kernel has reduced to below ∼2 pixels — by which point PSNR differences between views are already discriminative and the confidence signal is meaningful.

References

[1] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(4):376–380, 1991. 4

[2] Z. Wang, S. Wu, W. Xie, M. Chen, and V. A. Prisacariu. NeRF–: Neural Radiance Fields without Known Camera Parameters. https://arxiv.org/abs/2102.07064, 2021. arXiv:2102.07064. 3