DeSC: Learning Deep Semantic Descriptor for NeRF Registration

Sheldon Fung, Wei Pan, Kui Su, Hui Cui, Xinkui Zhao, Xuequan Lu†

***Abstract*—NeRF registration has gained increasing attention recently. While existing research demonstrates considerable potential for this task, most methods primarily focus on either global geometric or rendering photometric information during feature learning, overlooking the rich cross-modal information inherent in the NeRF embedding feature space. In this paper, we propose DeSC, a novel NeRF registration approach that leverages the rich cross-modal features from NeRF to learn robust semantic descriptors. In particular, we propose a Deep Semantic Aggregation module, which employs a weighted graph convolution network to capture high-frequency texture details in NeRF patches. This approach reveals the underlying semantics shared across different NeRFs of the same scene, thereby yielding more robust global feature descriptors that lead to better alignment accuracy and robustness. In addition, we design a density-aware photometric consistency loss that facilitates the learning of robust features. Extensive experimental results on Objaverse datasets demonstrate that our approach produces superior registration performance to state-of-the-art techniques.**

***Index Terms*—NeRF, Scene Registration, Deep Semantic Descriptor.**

I. Introduction

3D scene registration, which involves estimating the relative transformation between two given 3D scenes in arbitrary poses, is a fundamental problem in 3D computer graphics and vision. It facilitates various real-world applications such as autopilot, VR/AR, SLAM, etc. The registration task has been well-explored on common 3D data representations, such as point cloud or mesh.

Recent advances in the Neural Radiance Field (NeRF) [1] have brought a new perspective to the field. Research on Large-scale NeRFs [2], [3] shows the potential of concatenating NeRF blocks for city-scale representation. These methods assume all NeRF blocks are trained in a unified camera coordinate system. However, it is challenging to ensure the availability of relative camera poses during image capture in practice, as this requires high-precision multi-sensor fusion to track real-time trajectories and poses. NeRF2NeRF [4] pioneers the use of the surface field to perform registration on NeRFs in a global optimization manner. Nonetheless, it requires dull manual keypoint annotation that limits its practical use. DReg-NeRF [5] resorts to a deep-learning approach to directly predict the final set of correspondences with the aid of the Attention mechanism [6]. However, the point cloud sampled from NeRF is noisy and involves arbitrary artifacts due to inaccurate volume density estimations, especially in under-trained cases. Thus, the direct regressing correspondence is prone to failure in such scenarios. Alternatively, VF-NeRF [7] resorts to using the Normalizing Flow to select quality rays for registration optimization. Yet, it is vulnerable to textureless and highly symmetric cases since the optimization relies on photometric loss. Nevertheless, current methods only consider global geometric or rendering photometric information and neglect the rich cross-modal information that is exhibited in the intermediate view-independent embeddings of NeRF. These embeddings capture intrinsic scene properties, such as geometry and high-level semantics, and are less affected by view-dependent rendering inconsistencies or illumination artifacts.

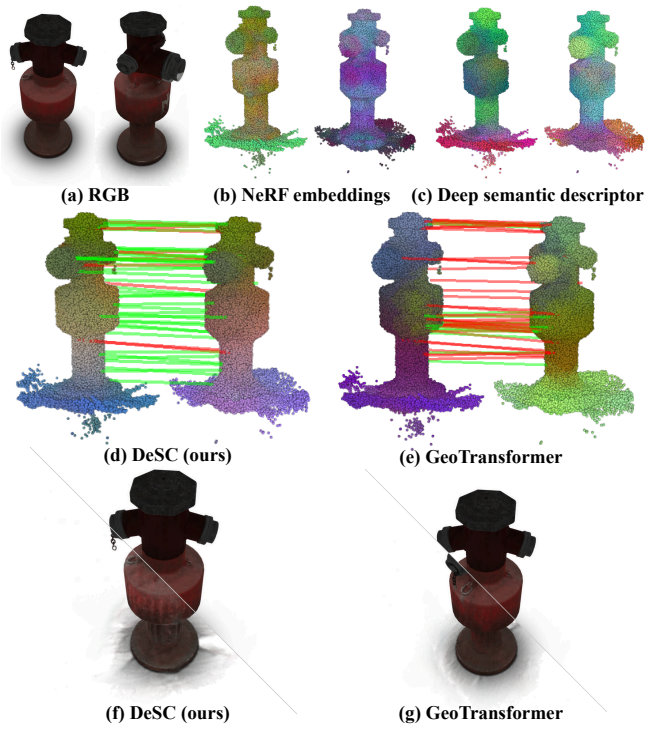

Fig. 1. Visualization of features in different stages. Despite showing strong correlations with color, NeRF embeddings exhibit cross-frame inconsistency (b). In contrast, our designed Deep Semantic Descriptor learns strong local view-independent textual semantic features and shows superior cross-frame consistency (c). This leads to more robust final features and yields better correspondences (d) for more precise NeRF alignment (f and g).

Sheldon Fung and Xuequan Lu are with The University of Western Australia. E-mail: sheldon.feng@research.uwa.edu.au, bruce.lu@uwa.edu.au.
Wei Pan is with OPTMV. E-mail: vpan@foxmail.com.
Kui Su is with Hangzhou City University. E-mail: suk@hzcu.edu.cn.
Hui Cui is with La Trobe University. E-mail: L.Cui@latrobe.edu.au.
Xinkui Zhao is with Zhejiang University. E-mail: zhaoxinkui@zju.edu.cn.
† corresponding authors.
Manuscript received August, 2024; revised April, 2025.

NeRF can be viewed as a 3D-2D cross-modal mapping function, where, given a viewing perspective and a point in the 3D space, NeRF outputs the corresponding color and density. In practice, NeRF decouples the prediction of color and density by first encoding the coordinate into a view-independent embedding with corresponding density, then a decoder is used to predict the color conditioned on the given perspective. We observe that the view-independent embedding exhibits rich cross-modal information as it encodes 3D coordinate and 2D color information. This inspires us to ask: how can we exploit such hybrid features to learn robust descriptors for the NeRF registration task? One naive solution is incorporating the point cloud registration pipeline and using the sampled view-dependent embeddings as inputs. However, the optimization in such a setting is non-trivial since 1) NeRFs are trained separately and do not share a universal embedding feature space, and 2) transforming the embedding feature into the same feature space compromises feature distinctiveness.

To this end, we propose DeSC, a novel NeRF registration approach that leverages the rich cross-modal features from NeRF to learn robust semantic descriptors. We design a Deep Semantic Aggregation module (DSAM) that leverages graph convolution [8] to strengthen local view-independent textual semantic features. Specifically, we sample point cloud generated from NeRF hierarchically using KPConv [9]. Given a sparse point, we construct a local patch by selecting k nearest dense points. Then we introduce a novel textual-aware graph convolution network on the patch to aggregate semantic feature descriptors. These features are subsequently fed to a Transformer [6] to learn global contextual information. Finally, the generated deep semantic feature descriptors can be used to predict the relative pose between the NeRF pair. Our technical contributions are summarized as follows:

- A cross-modal NeRF registration framework that leverages implicit color embeddings from NeRF to learn robust local feature descriptors.
- A transformation invariant deep semantic descriptor that captures local high-frequency implicit color features.
- a density-aware color consistency loss.

II. Related Works

Neural Radiance Field. In addition to widely adopted 3D scene representation techniques such as point clouds and meshes, Mildenhall *et al.* [1] introduce NeRF, an implicit 3D scene representation approach. NeRF utilizes differentiable volume rendering techniques and stores scene representations as the weights in a neural network through back-propagation. Despite its ability to synthesize novel views with high fidelity, the inefficiency during both training and testing phases poses significant limitations for real-world applications. Numerous works have been dedicated to addressing such issues by reducing redundant computation operations while maintaining the synthesize quality [10]–[22]. Liu *et al.* [10] utilizes voxel-bounded implicit fields and progressively prunes empty voxels during training to improve the computation efficiency. Instead of directly using rays for rendering, Mip-NeRF [16] improves fine details representation by alternatively rendering anti-aliased conical frustums. Barron *et al.* [16] extends Mip-NeRF to tackle artifacts presented in unbounded scenes. Instant-NGP [18] adopts a voxel-based occupancy grid to skip the redundant computation at empty space. Our work aims to align two trained NeRFs without known relative transformation.

NeRF Registration. Recent research [2], [3] shows the potential of dividing scenes into partitions to train multiple NeRFs for large-scale scene representation. However, these methods assume the relative transformation between two blocks is known, which is not always the case in real-world applications. iNeRF [23] estimates the relative pose between the image and NeRF, however, it does not directly estimate the relative pose between NeRF pair. Pioneer works [4], [24] incorporate global registration method (i.e. ICP [25]) and enforce surface consistency. These optimization-based methods require either fine initialization or human interactions. Reg-NF [26] introduces a bidirectional registration loss to avoid human annotation in [4]. Inspired by RegTr [27], DReg-NeRF [5] adopts a deep-learning approach and directly predicts the final set of correspondences between the point clouds sampled from the NeRF pair. VF-NeRF [7] proposes to use Normalizing Flows to generate camera perspective with high ViewShed Fields values to produce images for the registration task.

Point Cloud Registration. Correspondences-based methods, combined with outlier removal techniques such as RANSAC [28], are the most common approaches [29]–[45] for point cloud registration task. Early works employ convolution networks to extract local geometric features [29], [31]. With the Transformer [6] demonstrating superior performance across various tasks, DCP [32] makes the first attempt to adopt the Attention mechanism in the point cloud registration tasks. Predator [34] couples GCN [8] with cross attention to aggregate global feature representations. Recent studies seek to exploit 2D-3D cross-modal feature fusion [37], [38], [41]. Another research branch explores different strategies to enhance rotational variance [35], [36], [39], [44]. These state-of-the-art point cloud registration methods inspire our work. However, directly applying these methods to the NeRF registration task might result in a subpar performance since the point cloud sampled from NeRF tends to be noisy and contain artifacts. Our work takes an alternative approach and exploits the local implicit color embeddings from NeRF to learn robust semantic descriptors.

III. Motivation

Neural Radiance Fields (NeRFs) are powerful for 3D scene representation, but aligning independently trained NeRFs, known as NeRF registration, is challenging due to their implicit nature [5], [7]. Unlike explicit 3D data, such as point clouds, NeRFs encode scenes within neural networks, making standard registration methods less effective. Yet, NeRF registration is crucial due to the following aspects:

Scalability for large-scale scenes. In real-world scenarios, capturing and rendering large environments as a single NeRF can be computationally prohibitive and memory-intensive [2], [3], [5], [7]. Works such as Block-NeRF [2] and Mega-NeRF [3] have addressed this by partitioning large scenes into multiple, locally trained NeRF blocks. However, these blocks often

lack shared coordinate systems in practice and thus require registration to compose a globally coherent representation [5].

Handling pose uncertainty. Real-world data is frequently captured in GPS-denied or unconstrained environments where global camera poses are unavailable or noisy. As a result, NeRFs trained on different subsets of images will exist in arbitrary coordinate frames [5]. Registering these NeRFs into a unified frame becomes necessary for downstream tasks such as novel view synthesis, scene fusion, or long-term mapping [5], [7].

Efficiency and modularity. NeRF registration enables a modular reconstruction workflow: instead of re-training a monolithic NeRF whenever new image data becomes available, smaller NeRFs can be trained independently and later registered. This approach offers practical benefits in terms of training speed, incremental scene updates, and memory management.

In fact, NeRF registration presents unique challenges. Unlike explicit representations, NeRFs encode scenes implicitly in network weights, making standard registration methods less effective. Our proposed DeSC is designed specifically to operate in this implicit domain. Prior approaches like DReg-NeRF [5] often sample point clouds with RGB values from NeRFs and feed them into point cloud registration pipelines, such as RegTR [27] or GeoTR [35]. However, these sampled point clouds can be noisy and artifact-prone, especially in under-trained or low-overlap settings. In contrast, our method, DeSC, directly leverages NeRF's internal view-independent semantic embeddings, which are rich cross-modal features but are not available in point cloud data. We introduce a Deep Semantic Aggregation Module (DSAM) that aggregates these embeddings into robust local descriptors, which are then fused with geometric features and processed via Transformers [6] for matching. This design makes our approach more resilient to noise and overlap issues.

IV. Methodology

A. Problem Definition

Given two misaligned NeRF models F_N^X and F_N^Y, our objective is to find a set of correspondences $\mathbb{C}^* = \{(\mathbb{C}_{x_i}, \mathbb{C}_{y_i}) \mid \mathbb{C}_{x_i} \in \mathbb{R}^3, \mathbb{C}_{y_i} \in \mathbb{R}^3, i = 1, \ldots, t\}$ that lies within the bounding box of each NeRF model. Then by solving the following equation, we can obtain $\mathbf{R}^* \in SO(3)$ and $t^* \in \mathbb{R}^3$ that aligns F_N^X to F_N^Y:

$$\mathbf{R}^*, t^* = \min_{\mathbf{R},t} \sum_{(\mathbb{C}_{x_i}, \mathbb{C}_{y_i}) \in \mathbb{C}^*} \|\mathbf{R} \cdot \mathbb{C}_{x_i} + t - \mathbb{C}_{y_i}\|_2^2. \quad (1)$$

B. Querying NeRF

We consider NeRF an encoder-decoder network. Given a query point location $x \in \mathbb{R}^3$ within the bounding box, encoder network with parameter θ_e extracts volume density σ_x and view-independent embedding $e_x \in \mathbb{R}^{d_e}$:

$$\sigma_x, e_x = F_n(x; \theta_e). \quad (2)$$

In common practices [18], [23], the final color on the image is view-dependent and is derived from the integration along the sample rays. We follow [5] and compute the point color c_x at position x using the NeRF decoder with parameters θ_d:

$$c_x = \frac{1}{n} \sum_{i=0}^{n} F_n(r_i, e_x; \theta_d), \quad (3)$$

where n is the total image number used to train the NeRF. $r_i = (o_i, d_i)$ is the viewing perspective of the i-th image where o_i and d_i are the camera position and viewing direction, respectively. For the input NeRF N_X, we adopt the occupancy mask M_o from the occupancy grid in [18] to accelerate the sampling speed, and only sample the center of the non-empty voxel grid as point set $P \in \mathbb{R}^{\bar{m} \times 3}$. We then construct the point cloud X such that $X = \{p_i | p_i \in P, \sigma_{p_i} > \sigma_t\}$ and $X \in \mathbb{R}^{m \times 3}$, where σ_{p_i} can be computed using equation (2) and σ_t is a hyperparameter. Similarly, we construct point cloud $Y \in \mathbb{R}^{n \times 3}$ for the other input NeRF N_Y.

C. Deep Semantic Aggregation Module

NeRF serves as a cross-modal mapping function that maps 3D spatial coordinates to 2D visual observations. Although its view-independent latent embeddings capture rich semantic information across geometry and appearance, these embeddings are learned independently in each NeRF and remain fixed during registration, leading to significant distributional discrepancies across models.

To effectively extract semantic features from NeRFs, we propose the Deep Semantic Aggregation Module (DSAM) that leverages view-independent embeddings from NeRFs to learn robust semantic features. This module addresses the inconsistency of view-independent embeddings across separately trained NeRFs while still leveraging their rich semantic content.

Given a NeRF-derived point cloud X, we apply KP-FCNN [9] to sample X hierarchically and simultaneously extract multi-level geometry features. We use points and the corresponding features from the coarsest and the second dense layer, denoted as $\dot{\mathbf{X}} \in \mathbb{R}^{m' \times 3}$, $F^{\dot{X}} \in \mathbb{R}^{m' \times d_k}$, and $\hat{\mathbf{X}} \in \mathbb{R}^{n' \times 3}$, $F^{\hat{X}} \in \mathbb{R}^{n' \times d_k}$, respectively. d_k is the number of feature dimensions and $m' < n'$. Given a coarse point $x_i \in \dot{\mathbf{X}}$, we construct a patch p_i at dense points $\hat{\mathbf{X}}$:

$$p_i = \{x_j \in \hat{\mathbf{X}} \mid x_j \in K_n(x_i), \|x_i - x_j\|_2 < \tau\}, \quad (4)$$

where $K_n(\cdot)$ denotes the K-nearest neighbours and τ is a predefined radius threshold. To bridge the distribution gap between independently trained NeRFs, we enrich the view-independent embedding e_j with its corresponding density σ_i and color c_i, forming a photometrically conditioned embedding:

$$\hat{e}_i = e_i + \mathbf{S}_{4 \to d_e}([c_i; \sigma_i]), \quad (5)$$

where $[\cdot;\cdot]$ denotes concatenation operation and $\mathbf{S}_{4 \to d_e}(\cdot)$ is a deterministic projection function. We adopt sinusoidal positional encoding [6] as function $\mathbf{S}_{4 \to d_e}(\cdot)$ in this work.

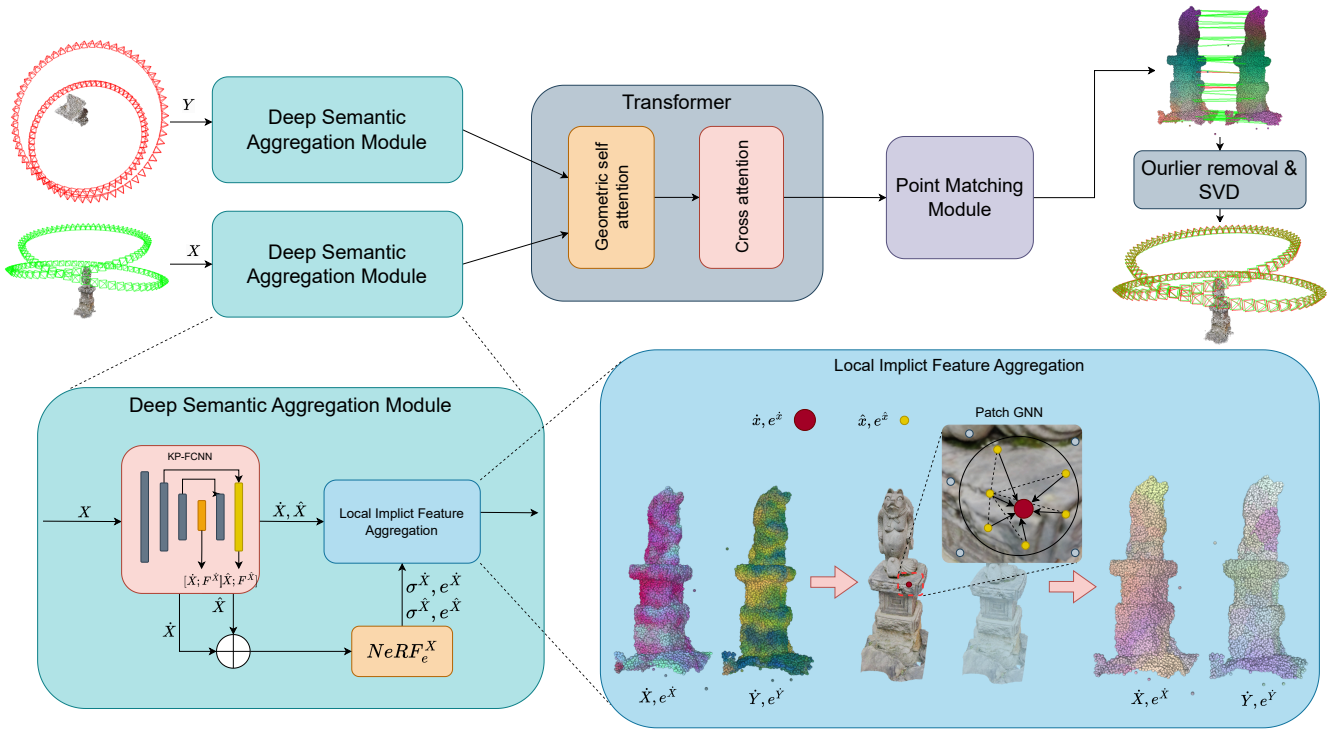

Fig. 2. An overview of our proposed DeSC. In our proposed Deep Semantic Aggregation module, FP-FCNN is employed to down-sample the point cloud hierarchically into multiple density levels and extract geometric features simultaneously. The second coarsest and the sparsest point clouds are used to query the respective NeRF to obtain view-independent embeddings. Our proposed texture-aware graph convolution network generates robust deep semantic descriptors from the NeRF view-independent embeddings which are then merged with geometric features.

By doing so, the benefits are twofold: 1) the output features are thus conditioned on the corresponding NeRF photometric information, which is shared by the NeRF pair N_X and N_Y, and 2) it strengthens the cross-frame mutual information that potentially stabilizes the learning process.

To capture local high-frequency textures, we introduce a novel texture-aware graph convolution network. For each patch p_i, we construct a local graph $\mathcal{G}_i = (p_i, \mathcal{E}_i)$, connecting neighbors based on proximity. We define the texture-aware edge convolution at layer $l+1$ as:

$$e_i^{(l+1)} = \max_{(x_j,x_k)\in\mathcal{E}_i} h_{\theta_l}([e_j^{(l)}; w_{jk}(e_k^{(l)} - e_j^{(l)})]), \tag{6}$$

where h_{θ_l} denotes convolution layer with weights θ_l. w_{jk} is the deviation weight that aims to capture the local high-frequency texture details that lie in the patch. It should emphasize points that are close to one another but exhibit distinct features. We thus formulate it as follows:

$$w_{jk} = exp(-\|x_j - x_i\|_2^2 - \alpha(\langle e_j, e_i\rangle + 1)), \tag{7}$$

where α is a hyperparameter. This design encourages the network to focus on feature contrasts in spatially close regions and is especially useful in capturing detailed textures and semantic boundaries. We concatenate all convolutional outputs and fuse them with the original geometric feature to form the final deep semantic descriptor e_i^x:

$$e_i^x = f_i + h_{\theta_g}([e_i^{(0)}; e_i^{(1)}; e_i^{(2)}; e_i^{(3)}]), \tag{8}$$

where h_{θ_g} denotes convolution layer with weights θ_g and $f_i \in F^{\dot{X}}$. Our graph-based aggregation strategy is crucial for capturing fine-grained local structures, especially in areas with high-frequency textures or complex geometry, ensuring robust feature consistency across source and target NeRFs (Figure 2). While simpler alternatives like average pooling or MLPs can be applied, they often overlook the spatial and contextual relationships between neighboring points. In contrast, our graph formulation enables adaptive, context-aware feature aggregation, yielding more expressive local descriptors and substantially enhancing registration robustness. The deep semantic descriptors are computed on sparsely sampled points to preserve distinctiveness. However, using only these sparse points for correspondence estimation may be insufficient. To address this, we adopt a coarse-to-fine registration framework [35] that progressively refines correspondences using dense geometric features (see Section IV-E).

D. Transformer

Inspired by previous works [35], [36], [39], [44], we employ the self-attention and cross-attention from the Tranformer [6] to learn robust features for matching. Self-attention aggregates the intra-frame global contextual information while cross-attention facilitates the inter-frame mutual information exchange as the yielded features are conditioned on the feature from the other frame.

Geometric self-attention. We follow [35] to enhance the intra-frame attention with transformation invariant geometric

encoding. We take the computation of point cloud X as an example. Given the input feature e_i^x yield from our Deep Semantic Aggregation module, the output of geometric self-attention $f_i^s \in \mathbb{R}^{m' \times d_t}$ is computed as the weighted sum of all projected input features:

$$f_i^s = \sum_{j=1}^{m'} w_{i,j}(e_i^x \mathbf{W}^v), \tag{9}$$

where $\mathbf{W}^v \in \mathbb{R}^{d_t \times d_t}$ is a learnable projection matrix. The attention weight $w_{i,j}$ is obtained through a row-wise softmax operation applied to the attention score $a_{i,j}^s$. And the attention score $a_{i,j}^s$ is obtained as follows:

$$a_{i,j}^s = \frac{(e_i^x \mathbf{W}^Q)(e_j^x \mathbf{W}^k + r_{i,j}\mathbf{W}^g)^T}{\sqrt{d_t}}, \tag{10}$$

where $\mathbf{W}^Q$, $\mathbf{W}^k$, and $\mathbf{W}^g \in \mathbb{R}^{d_t \times d_t}$ are learnable projection matrices. The geometric structure embedding $r_{i,j} \in \mathbb{R}^{d_t}$ comprises a pair-wise distance embedding and triple-wise angular embedding [35].

Feature-based cross-attention. We adopt vanilla cross-attention to allow mutual information exchange to aggregate cross-frame global contextual features. The final attention output f_i^c can be computed using equation (9). The attention score of cross-attention $a_{i,j}^c$ is computed as follows:

$$a_{i,j}^c = \frac{e_i^x \mathbf{W}^Q (e_j^y \mathbf{W}^k)^T}{\sqrt{d_t}}, \tag{11}$$

where $\mathbf{W}^Q$ and $\mathbf{W}^k \in \mathbb{R}^{d_t \times d_t}$ are learnable projection matrices. Note that the query e_i^x comes from NeRF N_X while the key e_j^y is from NeRF N_Y. The final output f_i^c is used in the subsequent matching procedure.

E. Matching

We adopt the coarse-to-fine matching strategy [35] to establish correspondences in a two-step manner. The coarse correspondences are established by selecting the top k entries of the Gaussian correlation matrix with normalized global features from the Transformer in section IV-D. To perform fine matching, the dense points are partitioned into patches using the points-to-node strategy [35]. Then for each established coarse correspondence, an optimal transport layer [46] is applied to the corresponding dense point patch partition to compute the cost matrix. Finally, the refined dense correspondences are generated with the mutual top k selection on the entries of cost matrices.

F. Supervision

Our loss function is defined $\mathcal{L} = \mathcal{L}_{oc} + \mathcal{L}_p + \mathcal{L}_d$. We follow [35] and adopt overlap-aware circle loss $\mathcal{L}_{oc}$ and point matching loss $\mathcal{L}_p$ to supervise sparse and dense features, respectively. Additionally, we introduce a novel density-aware photometric consistency loss $\mathcal{L}_d$ to improve the photometric consistency at the established correspondences.

Density-aware photometric consistency loss. Given two well-aligned NeRFs, the rendered pixels from the same camera perspective should be similar. With this prior knowledge, the photometric consistency loss aims to supervise the quality of the established correspondences by exploiting the similarity of the rendered color. However, the correspondences are established/refined using a non-differentiable coarse-to-fine strategy. Therefore, directly applying photometric consistency loss on fine correspondences is intractable, as dense point features only contain geometric information. Conversely, applying the loss directly at coarse correspondences is also unsuitable because they lack the necessary precision to accurately reflect photometric consistency. To tackle this dilemma, we instead exploit the patch-wise photometric information for supervision. Practically, we design the vague patch color $\bar{C}$ to capture the local patch photometric information which is formulated as follows:

$$\bar{c}_i = norm(\frac{1}{|p_i|}\sum_{j \in p_i} \sigma_j c_j), \tag{12}$$

where p_i is the patch of $\dot{x}_i$ computed through equation (4). c_j and σ_j are the color and density of $\hat{x}_j$, respectively. Note that the density σ_j is used as the weight of the corresponding color c_j. The reason is that the color information at low-density locations might be unreliable (e.g., the color of transparent glass). Then the density-aware photometric consistency loss $\mathcal{L}_d$ can be formulated as follows:

$$\mathcal{L}_d = \frac{1}{|\mathbb{C}_p|}\sum_{i,j \in \mathbb{C}_p} -log(\bar{c}_i \cdot \bar{c}_j^T), \tag{13}$$

where $\mathbb{C}_p$ is the predicted coarse correspondences. $\bar{c}_i$ and $\bar{c}_j$ are corresponding vague patch color from NeRF N_X and N_Y, respectively.

V. Experiments

A. Implementation

To evaluate our method, we implement it in PyTorch. We use the Adam optimizer during the training stage with an initial learning rate of 1e-4, a decay rate of 0.95 for every epoch, and a weight decay of 1e-6. All the training is conducted on an Nvidia A100 GPU. For the Objaverse dataset, we train the network for 20 epochs with a batch size of 1, which requires approximately 8 hours.

B. Evaluation on Object Scenes

Dataset. The Objaverse [47] is a text-to-3D dataset containing more than $800K$ objects. Building on the Objaverse dataset, the Chen *et al.* [5] construct a NeRF dataset by randomly selecting $30+$ categories that each contains $40-80$ objects/scenes, resulting in a total number of $1700+$ objects/scenes. For each sample, 120 images are rendered and split into 2 blocks by k-means. Then a randomly generated transformation is applied to the camera poses of the separated blocks to ensure the NeRFs are being trained in different camera coordinates. We select 44 unseen objects during training for the test. Each block of images is used to train a unique NeRF that is subsequently used to train and test our proposed approach.

TABLE I
REGISTRATION EVALUATION RESULTS ON THE OBJAVERSE DATASET. WE REPORT REGISTRATION RESULTS OF RRE, RTE, RMSE, AND RR MATRICES USING POINT CLOUD BASED METHODS (THE UPPER HALF) AND NERF-BASED METHODS (THE LOWER HALF). ADDITIONALLY, WE SHOW THE TOP 50% BEST RESULTS (THE LEFT MAIN COLUMN TITLED 50%).

Method	50%				100%			
	RRE (°)	RTE (cm)	RMSE	RR (%)	RRE (°)	RTE (cm)	RMSE	RR (%)
FGR [48]	10.77	9.42	10.01	82.6	13.21	11.24	12.47	79.3
CoFiNet [33]	7.55	6.81	7.97	85.7	11.07	9.09	10.10	82.4
predator [34]	8.07	6.73	8.02	85.5	12.11	8.82	10.72	82.1
Lepard [36]	5.62	3.88	4.42	95.3	10.88	3.71	7.43	92.6
GeoTR [35]	4.52	2.14	3.32	96.9	9.19	2.90	6.35	93.2
RegTR [27]	6.35	5.16	6.11	85.7	11.97	4.36	7.65	80.4
SIRA-PCR [40]	4.19	2.03	3.23	97.2	8.74	2.24	6.03	94.5
RoITr [39]	5.11	2.59	3.68	95.3	9.78	3.12	6.65	92.8
DReg-NeRF [5]	5.32	2.45	7.42	90.4	9.67	3.85	10.77	86.4
VF-NeRF [7]	1.96	1.94	2.57	97.1	6.77	2.14	5.34	92.1
Ours	**0.35**	**0.70**	**0.80**	**100.0**	**2.57**	**1.68**	**2.44**	**97.7**

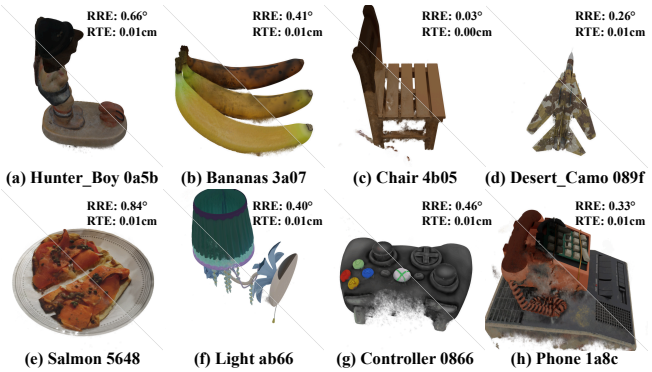

Fig. 3. Visualization of NeRF registration results. The rendered images of the aligned NeRFs are split by a diagonal line, where the top-right is the source NeRF and the bottom-left is the target NeRF. Note that despite the large alignment error for the high-symmetrical case (e.g. the cone in Figure (c)), our method can still generate visually accurate estimation.

Evaluation. Our method is evaluated on four metrics: 1) Relative Rotation Error (RRE), 2) Relative Translation Error (RTE), 3) Root Mean Squared Error (RMSE), and 4) Registration Recall (RR) RRE measures the rotational error in degrees and RTE measures the translation error in centimeters. RMSE measures the point-wise difference between the ground truth alignment and the transformation predicted by the model. RR calculates the fraction of point cloud pairs with a transformation RMSE less than $0.2m$, indicating the quality of the final alignment. See supplementary for details. NeRFs are individually trained on each sample, and the resulting models serve as inputs for registration. Performance is compared against state-of-the-art NeRF registration baselines: DReg-NeRF [5] and VF-NeRF [7]. We also compare with the state-of-the-art point cloud registration approaches: FGR [48], CoFiNet [33], Predator [34], Lepard [36], GeoTR [35], RegTR [27], SIRA-PCR [40], and RoITr [39]. Note that when comparing with these point cloud registration methods, we follow previous works [5], [7] to sample an occupancy grid [18] as the input point cloud.

Quantitative results. We report the registration evaluation results on the Objaverse dataset in table I. To highlight the effectiveness of our method, we include the comparisons with pointcloud-based methods (shown in the upper half in table I). Additionally, we report the top 50% of results based on the

TABLE II
COMPARISON OF THE ROBUSTNESS AGAINST DIFFERENT NOISE LEVELS ON VARIOUS METHODS. THE COLUMNS ON THE RIGHT SIDE DENOTE REGISTRATION RESULTS ON THE AMOUNT OF UNIFORM NOISE IN DIFFERENT SCALES (1%, 5%, 10% OF THE SCENE SIZE).

Method	RR (%)			
	0%	1%	5%	10%
FGR [48]	79.3	77.93	76.79	75.63
CoFiNet [33]	82.4	82.01	80.58	79.21
Predator [34]	82.1	81.81	80.31	79.89
Lepard [36]	92.6	92.18	91.51	90.94
GeoTR [35]	93.2	92.96	91.95	91.1
RegTR [27]	80.4	79.39	78.88	77.45
SIRA-PCR [40]	94.5	93.31	92.89	91.71
RoITr [39]	92.8	92.52	92.14	91.24
DReg-NeRF [5]	86.4	85.02	84.68	84.28
VF-NeRF [7]	92.1	91.31	90.18	88.93
Ours	**97.7**	**96.61**	**96.47**	**95.74**

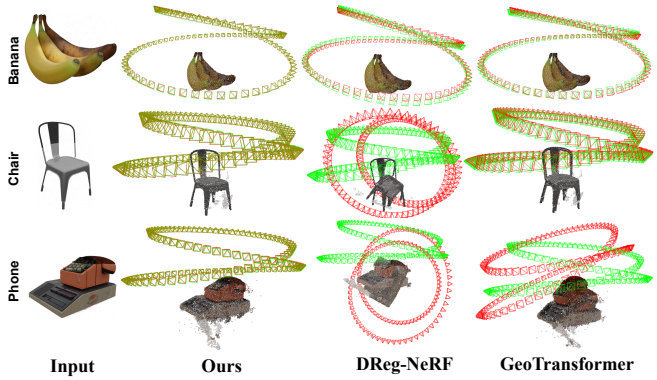

Fig. 4. Qualitative Registration Results and Comparisons. In addition to illustrating the predicted model alignments, we also visualize the camera poses associated with the input images for each respective NeRF model. For more intuitive results, we represent the camera poses from the input images of the source NeRF in red and those from the target NeRF in green. When applying our method, the precise alignment of these camera poses results in a yellow appearance, reflecting the mixture of red and green colors due to the accurate registration.

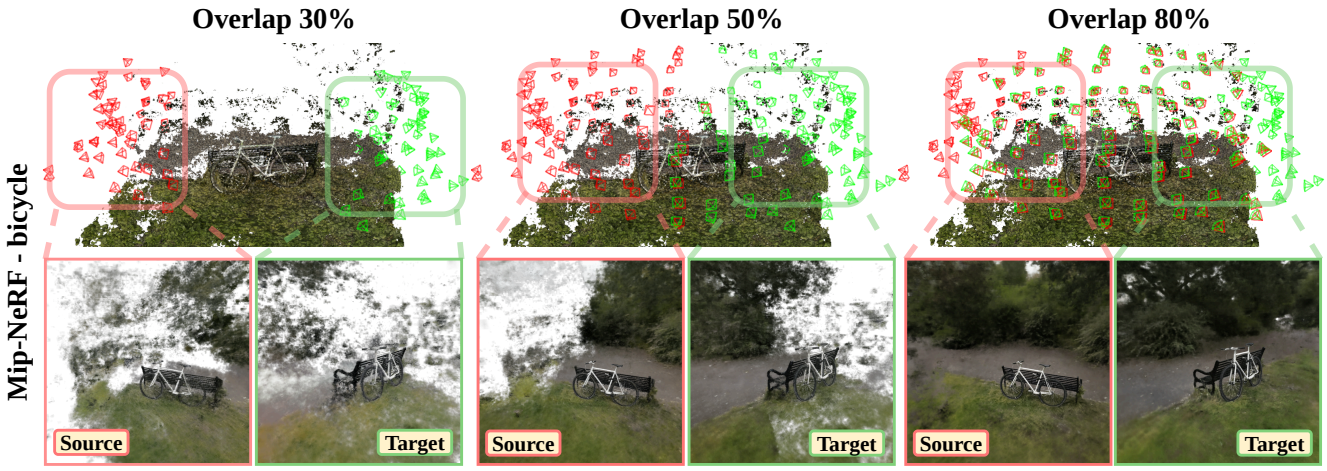

Fig. 5. An illustration of the settings of different overlap ratios. The left, middle, and right columns depict 30%, 50%, and 80% overlap ratios, respectively. The red and green frustums in the top row represent source and target images and their respective poses in the scenes. The bottom rows display the rendered images from the well-trained NeRF models from different perspectives.

RMSE metric for further analysis. Our method significantly surpasses the state-of-the-art techniques across all metrics. Specifically, our method outperforms the second-best NeRF registration method VF-NeRF [7] by 5.6% in RR, along with improvements of 4.2%, 0.46%, and 2.9% in terms of RRE, RTE, and RMSE, respectively. When compared to the point cloud-based method SIRA-PCR [40], our approach demonstrates a 3.2% improvement in RR, and respective gains of 6.17%, 0.56%, and 3.59% in terms of RRE, RTE, and RMSE. Similarly, for the top 50% results, our method outperforms VF-NeRF by 2.9% in RR, as well as by 1.61%, 1.24%, and 1.77% in RRE, RTE, and RMSE, respectively. Compared to SIRA-PCR [40], we achieve a 2.8% improvement in RR, along with gains of 3.84%, 1.33%, and 2.43% in RRE, RTE, and RMSE.

Qualitative results. We show comparisons of qualitative alignment results in figure 4. To better show the comparisons of alignment details, we add the camera pose visualization of the images used to train the corresponding NeRF model. We use red and green geometries to represent the image poses of the source and target models. Note that the camera geometries may appear yellow for well-aligned cases due to the color overlaps. We observe that in some cases, well-trained NeRFs are still noisy, i.e., having high-density values in empty space (see rows 2 and 3 in figure 4). We notice that this perturbation significantly influences the alignment results in some approaches (columns 3 and 4 in Figure 4). Nonetheless, our proposed method consistently achieves successful registration across all cases, demonstrating the robustness of the extracted features. Additionally, we also show the visualization of the aligned NeRF models in figure 3.

C. Evaluation on Large-scale Scenes

Datasets. To validate the generalizability of our method beyond synthetic datasets, we conduct comprehensive experiments on three real-world datasets with large-scale scenes: DTU [49], LLFF [50], and Mip-NeRF360 [51]. These datasets cover diverse scenarios, including controlled indoor captures, casually acquired forward-facing scenes, and unbounded 360-degree outdoor environments.

- **DTU dataset.** The DTU dataset [49] consists of 124 different *real-world scenes* and is divided into train (60%), test (30%), and validation (10%) sets. Each scene is augmented with arbitrary transformations and is separated into multiple sets of images with 30%, 50%, and 80% overlap ratios (See Figure 5), making a total of 744, 372, and 124 for training, testing, and validation sets, respectively.
- **LLFF dataset.** The LLFF dataset [50] consists of 8 *real-world large-scale scenes*. We divide and augment the dataset similar to the DTU dataset mentioned above, and the resulting sample numbers are 480, 240, and 80 for training, testing, and validation sets, respectively.
- **Mip-NeRF360 dataset.** The Mip-NeRF360 dataset [51] consists of 9 *real-world large-scale scenes*. We divide and augment the dataset similar to the DTU dataset mentioned above, and the resulting sample numbers are 540, 270, and 90 for training, testing, and validation sets, respectively.

Evaluation. The evaluation metrics used are Relative Rotation Error (RRE) in degrees, Relative Translation Error (RTE) in centimeters, Root Mean Squared Error (RMSE), and Registration Recall (RR) in percentage. NeRFs are individually trained on these subsets, and the resulting models serve as inputs for registration. Performance is compared against state-of-the-art NeRF registration baselines: DReg-NeRF [5] and VF-NeRF [7]. We also compare with the state-of-the-art point cloud registration approaches: RegTr [27], SIRA-PCR [40], and RoITr [39]. Note that when comparing with these point cloud registration methods, we follow previous works [5], [7] to sample an occupancy grid [18] as the input point cloud.

Quantitative results. The results on the DTU dataset (Table III) show that our method outperforms existing approaches across most evaluation metrics. Our method achieves the lowest RTE and RMSE across all overlap settings, demonstrating

TABLE III
REGISTRATION PERFORMANCE COMPARISONS ON THE DTU DATASET.

Method	30%				50%				80%			
	RRE (°)	RTE (cm)	RMSE	RR (%)	RRE (°)	RTE (cm)	RMSE	RR (%)	RRE (°)	RTE (cm)	RMSE	RR (%)
RegTR [27]	8.50	8.11	16.24	75.5	5.15	4.15	10.59	86.2	3.42	2.73	6.85	89.6
SIRA-PCR [40]	7.06	7.43	15.10	79.9	4.93	4.53	9.31	84.2	2.44	1.77	4.96	93.9
RoITr [39]	7.17	6.36	14.51	75.7	4.66	3.10	8.21	89.9	2.50	1.79	4.49	95.7
DReg-NeRF [5]	5.61	5.23	10.75	85.1	2.87	2.13	4.35	93.4	**1.05**	1.96	2.14	97.7
VF-NeRF [7]	**3.82**	4.30	8.41	86.6	3.34	2.20	4.99	93.0	1.77	**1.09**	2.92	97.2
Ours	3.90	**3.46**	**7.98**	**90.6**	**2.58**	**1.45**	**3.52**	**95.3**	1.21	1.82	**2.01**	**97.9**

TABLE IV
REGISTRATION PERFORMANCE COMPARISONS ON THE LLFF DATASET.

Method	30%				50%				80%			
	RRE (°)	RTE (cm)	RMSE	RR (%)	RRE (°)	RTE (cm)	RMSE	RR (%)	RRE (°)	RTE (cm)	RMSE	RR (%)
RegTR [27]	9.55	7.01	17.08	70.7	5.75	4.85	11.29	80.8	2.86	2.32	4.47	93.6
SIRA-PCR [40]	7.53	6.31	13.75	75.7	4.99	2.97	9.02	86.5	2.95	2.44	5.14	92.1
RoITr [39]	7.11	5.34	13.32	78.5	3.64	3.27	6.91	87.3	2.47	1.87	3.75	94.8
DReg-NeRF [5]	5.73	4.45	9.48	84.5	2.87	2.51	5.40	91.1	**1.55**	1.32	2.32	98.0
VF-NeRF [7]	4.82	3.50	7.92	88.5	2.51	1.69	3.96	92.6	1.60	1.12	2.11	97.9
Ours	**3.69**	**2.70**	**6.08**	**91.0**	**1.98**	**1.22**	**2.81**	**94.2**	2.04	**0.91**	**1.92**	**98.1**

TABLE V
REGISTRATION PERFORMANCE COMPARISONS ON THE MIP-NERF360 DATASET.

Method	30%				50%				80%			
	RRE (°)	RTE (cm)	RMSE	RR (%)	RRE (°)	RTE (cm)	RMSE	RR (%)	RRE (°)	RTE (cm)	RMSE	RR (%)
RegTR [27]	13.51	8.05	23.50	71.5	7.10	4.77	13.82	83.7	3.24	2.67	5.97	91.6
SIRA-PCR [40]	13.24	7.02	21.03	74.9	5.73	5.05	13.09	80.9	3.91	2.65	7.37	87.1
RoITr [39]	11.66	6.07	20.38	79.4	5.18	3.95	12.29	85.5	3.07	2.31	6.71	89.1
DReg-NeRF [5]	11.16	4.71	18.78	80.9	4.24	3.09	9.65	86.3	2.59	1.88	4.93	92.5
VF-NeRF [7]	10.31	4.00	16.63	84.8	3.68	2.55	8.23	89.2	1.80	1.59	3.75	93.9
Ours	**9.20**	**3.09**	**13.78**	**88.1**	**3.21**	**2.11**	**7.12**	**90.9**	**1.52**	**1.28**	**2.95**	**95.3**

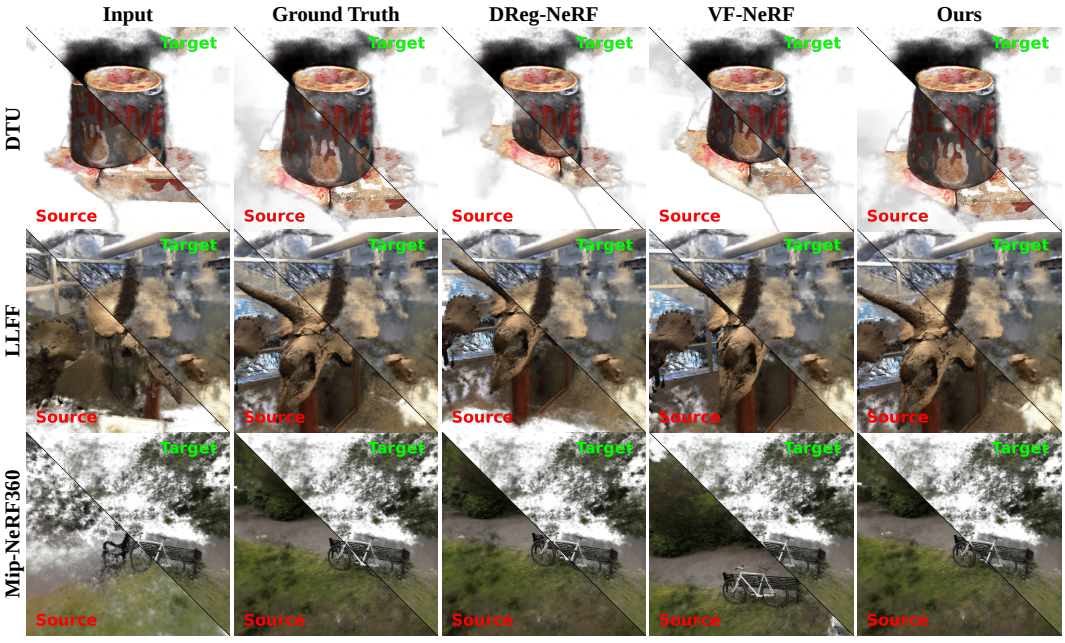

Fig. 6. The comparisons of qualitative results across different methods. We show the visual comparison of samples with a 30% overlap ratio from DTU, LLFF, and Mip-NeRF360 datasets.

superior robustness in estimating accurate transformations. The Registration Recall (RR) is consistently the highest,

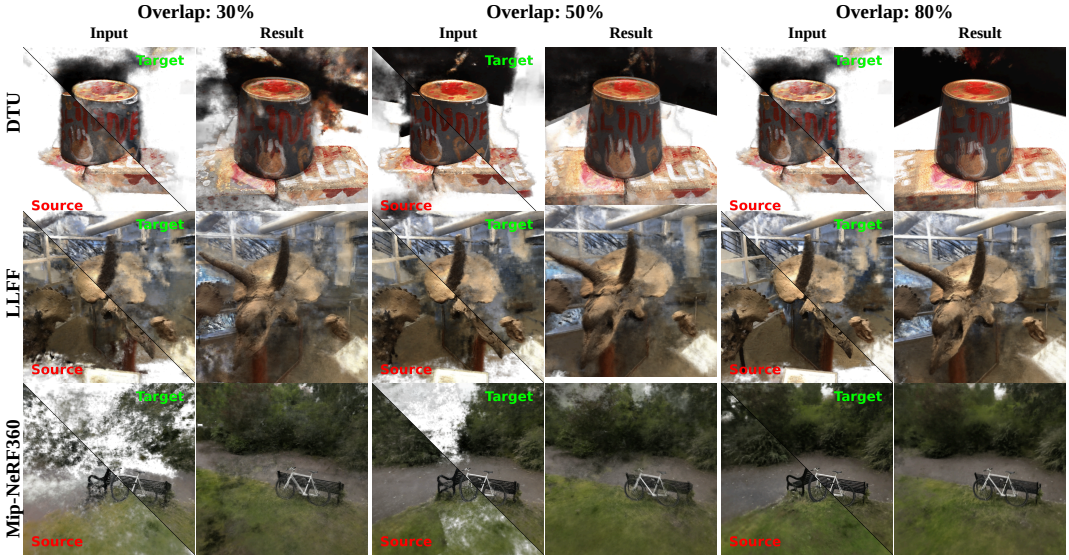

Fig. 7. The qualitative results of different overlap ratios for our proposed method. The left, middle, and right columns are samples with 30%, 50%, and 80% overlap ratios, respectively.

especially at 30% overlap, where our method outperforms the second-best approach (VF-NeRF [7]) by 4.0% (90.6% vs. 86.6%). While VF-NeRF achieves slightly lower RRE in some settings, our method maintains a strong balance between rotational and translational errors, leading to a more reliable registration performance.

The results in Table IV indicate that our method continues to perform favorably compared to other methods. Our approach consistently achieves the best RMSE and RTE, reaffirming its robustness in aligning images, even in large-scale real-world scenes. At 30% overlap, our method outperforms all baselines in RR (91.0%), which is a 2.5% improvement over VF-NeRF [7] (88.5%). While VF-NeRF has a slightly lower RRE at 80% overlap, our method demonstrates an overall balanced performance across all overlap levels, ensuring both accuracy and reliability.

The results in Table V further validate the effectiveness of our method in more complex and large-scale 360-degree scenes. Our method consistently achieves the best performance across all metrics, significantly outperforming previous state-of-the-art approaches. Compared to VF-NeRF [7], our method improves RTE ($3.09cm$ vs. $4.00cm$ at 30% overlap) and RR (88.1% vs. 84.8%), demonstrating its strong generalization ability. The improvement is particularly notable in low-overlap cases, where other methods struggle with larger errors, while our method maintains stable and accurate registration.

Qualitative comparisons. We visualize the qualitative result comparisons across different methods in Figure 6 in low overlap scenes (30% overlaps). In particular, the scene from Mip-NeRF360 V is the largest in scale among the scene-level datasets. The low overlap ratio causes massive empty space with noise artifacts. Nonetheless, our method is still able to extract effective features that lead to accurate NeRF registration, while other methods fail. We visualize the qualitative result with different overlap ratios for our method in Figure 7. It shows that our proposed method can achieve accurate NeRF registration results under extreme low overlap ratios, showcasing the robustness of the method.

D. Ablation

TABLE VI
ABLATION RESULTS SHOWCASING THE IMPACT OF THE DEEP SEMANTIC AGGREGATION MODULE (DSAM), THE PHOTOMETRIC ENCODING (PE), THE LOCAL HIGH-FREQUENCY TEXTURE WEIGHTS (LHTW), AND DENSITY-AWARE PHOTOMETRIC CONSISTENCY LOSS (DPCL) ON THE NERF REGISTRATION TASK FOR THE OBJAVERSE DATASET. THE "×" SIGN REPRESENTS THE CORRESPONDING MODULE BEING REMOVED.

DSAM	PE	LHTW	DPCL	RRE (°)	RTE (cm)	RMSE	RR (%)
×	×	×	×	9.19	2.90	0.063	93.2
✓	×	×	×	7.57	2.83	0.061	93.9
✓	✓	×	×	6.11	2.44	0.055	95.1
✓	✓	✓	×	3.90	1.96	0.039	96.9
✓	✓	✓	✓	**2.57**	**1.70**	**0.024**	**97.7**

We conduct extensive ablation experiments to study the significance of our proposed Deep Semantic Aggregation Module (DSAM) and Density-aware photometric consistency loss (DPCL). Particularly, we also perform ablation experiments on the photometric encoding (PE) and the local high-frequency texture weights (LHTW) within the proposed DSAM. Experiments results are summarized in table VI. Note that when PE is removed, we directly use NeRF embedding as input features. Similarly, we adopt vanilla edge convolution [8] when the LHTW is removed. In general, when both DSAM and DPCL are removed (row 1), the pipeline degenerates to a pure point cloud registration method similar to [35]. This setting is utilized as our baseline to estimate the effectiveness of our proposed modules. We summarize the ablation study by answering the following questions:

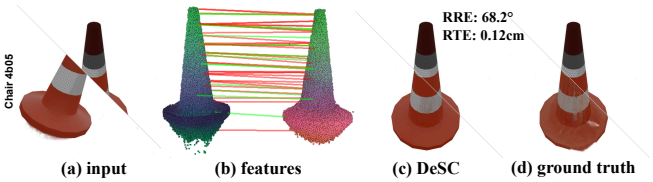

Fig. 8. Visualization of a failure case. Our method is able to generate visually accurate alignment despite the large transformation error due to the symmetrical structure.

How effective are the proposed modules? When all modules are removed, the overall performance drops significantly. Specifically, RR drops dramatically from 97.7% to 93.2%. For alignment details, RMSE increases by a large margin of 0.039. Meanwhile, both RRE and RTE saw increasing trends, with RRE rising from 2.57 to 9.19 degrees and RTE increasing from 1.70 cm to 2.90 cm. The results demonstrate the effectiveness of our proposed modules.

Is the PE and the LHTW significant? The overall performance saw a notable descent across all metrics by removing both PE and the LHTW in the DSAM (row 2). Compared to the default setting (row 5), RR declined by 3.8%. Similarly, RMSE increased remarkably by 0.037, indicating a deterioration in alignment accuracy. Additionally, both RRE and RTE exhibited increasing trends, further highlighting the importance of PE and LHTW in maintaining high performance. Their individual effects are also evident, as either of their presence improves the overall performance (row 3 and row 4). Notably, the LHTW has a more significant effect, as its presence results in a higher RR increase of 0.6%. A similar trend is also evident in RMSE.

Is the DPCL helpful? When DPCL is included (row 5), the overall performance shows a notable improvement across all metrics compared to the opposite (row 4). Specifically, RRE and RTE both decrease significantly, RMSE is reduced, and RR increases from 96.9% to 97.7%. This indicates that DPCL plays a crucial role in enhancing the alignment accuracy and robustness of the registration process.

VI. Limitation

Our method is designed specifically for NeRF-based representations. It makes use of NeRF's implicit embedding features through the Deep Semantic Aggregation module to extract local high-frequency color details. As a result, it does not directly apply to explicit 3D representations like 3D Gaussian Splatting (3DGS), voxel grids, or hash-based methods. However, since 3DGS includes rich per-point attributes like 3D Gaussians and spherical harmonics, adapting our approach to 3DGS could be a valuable direction for future research.

Another limitation is with highly symmetrical objects or scenes. For example, as shown in Figure 8, the traffic cone has strong symmetry in both shape and texture. This causes our extracted features to also appear symmetric, which can lead to incorrect matches during registration (see Figure 8(b)). We plan to investigate this issue further in future work.

VII. Conclusion

In this paper, we presented DeSC, a novel NeRF registration approach that leverages the rich cross-modal features from NeRF to learn robust semantic descriptors. Specifically, we designed a deep semantic aggregation module that learns local view-independent textual semantic features. It leverages a photometric encoding to strengthen the cross-frame mutual information and a novel texture-aware graph convolution network to capture local high-frequency textures. In addition, we design a density-aware photometric consistency loss to facilitate the learning of robust features. Our experiment results showcase the superiority of our model.

References

[1] B. Mildenhall, P. P. Srinivasan, M. Tancik, J. T. Barron, R. Ramamoorthi, and R. Ng, "Nerf: Representing scenes as neural radiance fields for view synthesis," *Communications of the ACM*, vol. 65, no. 1, pp. 99–106, 2021.

[2] M. Tancik, V. Casser, X. Yan, S. Pradhan, B. Mildenhall, P. P. Srinivasan, J. T. Barron, and H. Kretzschmar, "Block-nerf: Scalable large scene neural view synthesis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 8248–8258.

[3] H. Turki, D. Ramanan, and M. Satyanarayanan, "Mega-nerf: Scalable construction of large-scale nerfs for virtual fly-throughs," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 922–12 931.

[4] L. Goli, D. Rebain, S. Sabour, A. Garg, and A. Tagliasacchi, "nerf2nerf: Pairwise registration of neural radiance fields," in *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2023, pp. 9354–9361.

[5] Y. Chen and G. H. Lee, "Dreg-nerf: Deep registration for neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 22 703–22 713.

[6] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[7] L. Segre and S. Avidan, "Vf-nerf: Viewshed fields for rigid nerf registration," in *European Conference on Computer Vision*. Springer, 2024, pp. 164–181.

[8] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic graph cnn for learning on point clouds," *ACM Transactions on Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[9] H. Thomas, C. R. Qi, J.-E. Deschaud, B. Marcotegui, F. Goulette, and L. J. Guibas, "Kpconv: Flexible and deformable convolution for point clouds," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 6411–6420.

[10] L. Liu, J. Gu, K. Zaw Lin, T.-S. Chua, and C. Theobalt, "Neural sparse voxel fields," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15 651–15 663, 2020.

[11] S. J. Garbin, M. Kowalski, M. Johnson, J. Shotton, and J. Valentin, "Fastnerf: High-fidelity neural rendering at 200fps," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 14 346–14 355.

[12] J. T. Barron, B. Mildenhall, M. Tancik, P. Hedman, R. Martin-Brualla, and P. P. Srinivasan, "Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5855–5864.

[13] S. Vora, N. Radwan, K. Greff, H. Meyer, K. Genova, M. S. Sajjadi, E. Pot, A. Tagliasacchi, and D. Duckworth, "Nesf: Neural semantic fields for generalizable semantic segmentation of 3d scenes," *arXiv preprint arXiv:2111.13260*, 2021.

[14] S. Zhi, T. Laidlow, S. Leutenegger, and A. J. Davison, "In-place scene labelling and understanding with implicit scene representation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 838–15 847.

[15] A. Yu, R. Li, M. Tancik, H. Li, R. Ng, and A. Kanazawa, "Plenoctrees for real-time rendering of neural radiance fields," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 5752–5761.

[16] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5470–5479.
[17] K. Deng, A. Liu, J.-Y. Zhu, and D. Ramanan, "Depth-supervised nerf: Fewer views and faster training for free," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 12 882–12 891.
[18] T. Müller, A. Evans, C. Schied, and A. Keller, "Instant neural graphics primitives with a multiresolution hash encoding," *ACM transactions on graphics (TOG)*, vol. 41, no. 4, pp. 1–15, 2022.
[19] X. Fu, S. Zhang, T. Chen, Y. Lu, L. Zhu, X. Zhou, A. Geiger, and Y. Liao, "Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022, pp. 1–11.
[20] V. Tschernezki, I. Laina, D. Larlus, and A. Vedaldi, "Neural feature fusion fields: 3d distillation of self-supervised 2d image representations," in *2022 International Conference on 3D Vision (3DV)*. IEEE, 2022, pp. 443–453.
[21] S. Fridovich-Keil, A. Yu, M. Tancik, Q. Chen, B. Recht, and A. Kanazawa, "Plenoxels: Radiance fields without neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 5501–5510.
[22] A. Chen, Z. Xu, A. Geiger, J. Yu, and H. Su, "Tensorf: Tensorial radiance fields," in *European Conference on Computer Vision*. Springer, 2022, pp. 333–350.
[23] L. Yen-Chen, P. Florence, J. T. Barron, A. Rodriguez, P. Isola, and T.-Y. Lin, "inerf: Inverting neural radiance fields for pose estimation. in 2021 ieee," in *RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1323–1330.
[24] C. Peat, O. Batchelor, R. Green, and J. Atlas, "Zero nerf: Registration with zero overlap," *CoRR, abs/2211.12544*, 2022.
[25] P. J. Besl and N. D. McKay, "Method for registration of 3-d shapes," in *Sensor fusion IV: control paradigms and data structures*, vol. 1611. Spie, 1992, pp. 586–606.
[26] S. Hausler, D. Hall, S. Mahendren, and P. Moghadam, "Reg-nf: Efficient registration of implicit surfaces within neural fields," *arXiv preprint arXiv:2402.09722*, 2024.
[27] Z. J. Yew and G. H. Lee, "Regtr: End-to-end point cloud correspondences with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 6677–6686.
[28] R. Schnabel, R. Wahl, and R. Klein, "Efficient ransac for point-cloud shape detection," in *Computer graphics forum*, vol. 26, no. 2. Wiley Online Library, 2007, pp. 214–226.
[29] C. Choy, J. Park, and V. Koltun, "Fully convolutional geometric features," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 8958–8966.
[30] R. B. Rusu, N. Blodow, and M. Beetz, "Fast point feature histograms (fpfh) for 3d registration," in *2009 IEEE international conference on robotics and automation*. IEEE, 2009, pp. 3212–3217.
[31] X. Bai, Z. Luo, L. Zhou, H. Fu, L. Quan, and C.-L. Tai, "D3feat: Joint learning of dense detection and description of 3d local features," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 6359–6367.
[32] Y. Wang and J. M. Solomon, "Deep closest point: Learning representations for point cloud registration," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 3523–3532.
[33] H. Yu, F. Li, M. Saleh, B. Busam, and S. Ilic, "Cofinet: Reliable coarse-to-fine correspondences for robust pointcloud registration," *Advances in Neural Information Processing Systems*, vol. 34, pp. 23 872–23 884, 2021.
[34] S. Huang, Z. Gojcic, M. Usvyatsov, A. Wieser, and K. Schindler, "Predator: Registration of 3d point clouds with low overlap," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2021, pp. 4267–4276.
[35] Z. Qin, H. Yu, C. Wang, Y. Guo, Y. Peng, and K. Xu, "Geometric transformer for fast and robust point cloud registration," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 11 143–11 152.
[36] Y. Li and T. Harada, "Lepard: Learning partial point cloud matching in rigid and deformable scenes," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 5554–5564.
[37] Y. Zhang, J. Yu, X. Huang, W. Zhou, and J. Hou, "Pcr-cg: Point cloud registration via deep explicit color and geometry," in *European Conference on Computer Vision*. Springer, 2022, pp. 443–459.
[38] M. Yuan, K. Fu, Z. Li, Y. Meng, and M. Wang, "Pointmbf: A multi-scale bidirectional fusion network for unsupervised rgb-d point cloud registration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 17 694–17 705.
[39] H. Yu, Z. Qin, J. Hou, M. Saleh, D. Li, B. Busam, and S. Ilic, "Rotation-invariant transformer for point cloud matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5384–5393.
[40] S. Chen, H. Xu, R. Li, G. Liu, C.-W. Fu, and S. Liu, "Sira-pcr: Sim-to-real adaptation for 3d point cloud registration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 14 394–14 405.
[41] J. Yu, L. Ren, Y. Zhang, W. Zhou, L. Lin, and G. Dai, "Peal: Prior-embedded explicit attention learning for low-overlap point cloud registration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 702–17 711.
[42] X. Lu, H. Chen, S.-K. Yeung, Z. Deng, and W. Chen, "Unsupervised articulated skeleton extraction from point set sequences captured by a single depth camera," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.
[43] X. Lu, Z. Deng, J. Luo, W. Chen, S.-K. Yeung, and Y. He, "3d articulated skeleton extraction using a single consumer-grade depth camera," *Computer Vision and Image Understanding*, vol. 188, p. 102792, 2019.
[44] S. Fung, W. Pan, X. Liu, J. Yearwood, R. Dazeley, and X. Lu, "Topformer: Topology-aware transformer for point cloud registration," in *International Conference on Computational Visual Media*. Springer, 2024, pp. 112–128.
[45] S. Fung, X. Lu, D. de Silva Edirimuni, W. Pan, X. Liu, and H. Li, "Semreg: Semantics constrained point cloud registration," in *European Conference on Computer Vision*. Springer, 2024.
[46] P.-E. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich, "Superglue: Learning feature matching with graph neural networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 4938–4947.
[47] M. Deitke, D. Schwenk, J. Salvador, L. Weihs, O. Michel, E. VanderBilt, L. Schmidt, K. Ehsani, A. Kembhavi, and A. Farhadi, "Objaverse: A universe of annotated 3d objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 13 142–13 153.
[48] Q.-Y. Zhou, J. Park, and V. Koltun, "Fast global registration," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. Springer, 2016, pp. 766–782.
[49] H. Aanæs, R. R. Jensen, G. Vogiatzis, E. Tola, and A. B. Dahl, "Large-scale data for multiple-view stereopsis," *International Journal of Computer Vision*, pp. 1–16, 2016.
[50] B. Mildenhall, P. P. Srinivasan, R. Ortiz-Cayon, N. K. Kalantari, R. Ramamoorthi, R. Ng, and A. Kar, "Local light field fusion: Practical view synthesis with prescriptive sampling guidelines," *ACM Transactions on Graphics (TOG)*, 2019.
[51] J. T. Barron, B. Mildenhall, D. Verbin, P. P. Srinivasan, and P. Hedman, "Mip-nerf 360: Unbounded anti-aliased neural radiance fields," *CVPR*, 2022.

Sheldon Fung received a Bachelor's degree from Tiangong University, China. He is currently a Ph.D. candidate at the University of Western Australia under the supervision of Prof. Xuequan Lu. His research focuses specifically on the field of computer vision. His research interests encompass a wide range of topics within this field, including point cloud processing, deep learning algorithms, and computer graphics.

Wei Pan is currently working in OPT Machine Vision Corp. as a research lead in machine vision algorithm and system development. Prior to that, he worked as a research fellow at Shenzhen University and South China University of Technol-

ogy after he received his Ph.D. degree from Singapore University of Technology and Design. His research interests include 3D imaging, 3D Data processing, computer vision, machine learning, computer graphics and computer-aided design.

Kui Su has been a Research Faculty member at Hangzhou City University since 2022. He received a Ph.D degree at the School of Computer Science and Technology, Zhejiang University in 2017, and worked as a postdoctoral researcher in Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences(2018-2021). He once worked at companies including Huawei, Sangfor Technologies, and Alibaba DAMO Academy. His current research interests include Cloud Computing and Artificial Intelligence Computing.

Hui Cui is a Senior Lecturer at the Department of CS&IT, La Trobe University. From 2016 to 2018, she was a postdoctoral researcher at the University of Sydney (USYD), Sydney, Australia. She received MPhil and PhD degrees from USYD in 2013 and 2016 respectively, under the supervision of Prof David Feng, A/Prof Xiuying Wang, and Clinical Professor Michael Fulham.

Xinkui Zhao is a ZJU 100-Young Professor at Zhejiang University, China. His research interests primarily focus on AI4System, Data Mining, and Artificial Intelligence. He has led the development of multiple large-scale, enterprise-level cloud-native platforms and has authored over 30 academic papers published in top-tier journals and conferences, including ASPLOS, DAC, WWW, SCIS, TPDS, TSC, ICWS, ICSOC, and more.

Xuequan Lu (Senior Member, IEEE) is a Senior Lecturer at the Department of Computer Science and Software Engineering, The University of Western Australia (UWA), Australia. He spent more than two years as a Research Fellow in Singapore. Prior to that, he earned his PhD at Zhejiang University. His research interests mainly fall into visual computing, e.g., 3D vision/graphics, VR/AR, 2D image processing, and digital health. More information can be found at http://www.xuequanlu.com.