Learning 3D Scene Analogies with Neural Contextual Scene Maps

Junho Kim[1], Gwangtak Bae[1], Eun Sun Lee[1], and Young Min Kim[1, 2]
[1] Dept. of Electrical and Computer Engineering, Seoul National University
[2] Interdisciplinary Program in Artificial Intelligence and INMC, Seoul National University

(a) Scene Regions with Similar Spatial Context

Object Placements

(b) Dense Spatial Maps from **3D Scene Analogies**

Motion Trajectories

(c) Applications in Fine-Grained 3D Transfer

Fig. 1: Overview of the 3D scene analogy task. (a) Given two scenes with regions possibly having similar contexts, (b) the 3D scene analogy task aims to find a dense 3D mapping between the corresponding regions. (c) The estimated maps can then be used for applications such as object placement or motion trajectory transfer.

***Abstract*—Understanding scene contexts is crucial for machines to perform tasks and adapt prior knowledge in unseen or noisy 3D environments. As data-driven learning is intractable to comprehensively encapsulate diverse ranges of layouts and open spaces, we propose teaching machines to identify relational commonalities in 3D spaces. Instead of focusing on point-wise or object-wise representations, we introduce 3D scene analogies, which are smooth maps between 3D scene regions that align spatial relationships. Unlike well-studied single instance-level maps, these scene-level maps smoothly link large scene regions, potentially enabling unique applications in trajectory transfer in AR/VR, long demonstration transfer for imitation learning, and context-aware object rearrangement. To find 3D scene analogies, we propose neural contextual scene maps, which extract descriptor fields summarizing semantic and geometric contexts, and holistically align them in a coarse-to-fine manner for map estimation. This approach reduces reliance on individual feature points, making it robust to input noise or shape variations. Experiments demonstrate the effectiveness of our approach in identifying scene analogies and transferring trajectories or object placements in diverse indoor scenes, indicating its potential for robotics and AR/VR applications. Project page including the code is available through this link: https://82magnolia.github.io/3d_scene_analogies/.**

I. Introduction

The 3D world is rich in contextual information, shaped by the interplay of object placements and surrounding open spaces [66, 5]. The function of an object is often flexible, shifting according to its location and spatial relationship to nearby elements; a table might serve as a TV stand in one context or as a tea table beside a sofa in another. Capturing these nuanced, high-dimensional relationships is challenging. Decades of research in cognitive psychology [27, 52, 28, 29, 31, 73] suggest that humans rely on analogical reasoning to relate familiar scenes from past experiences to new observations. In Figure 1, humans can intuitively relate areas near a sofa-and-table setup in one room to similar areas in another, yet enabling machines to perform this mapping is far from straightforward. To achieve this, one must transfer not only the positions of objects but also their surrounding context which cannot be done through simple object or point-wise matching. How can we formulate this problem and extract generalizable representations that encode intricate object relationships and spatial context?

To address these challenges, we propose the **3D scene analogy task** of estimating a *dense map* between scenes that share similar contexts, as shown in Figure 1. This task demands a smooth map that preserves spatial coherence, allowing consistent relationships across mapped regions without abrupt transitions. By capturing both individual object placements and their surrounding context, the mapping enables transferring spatial arrangements between scenes in a structure-aware manner.

This contrasts with conventional feature matching from vision foundation models [7, 70, 74] or 3D keypoints [15, 101], which are often computationally costly or lack scalability for fine-grained scene mapping. Moreover, these features struggle to capture semantic relationships or nuanced contextual cues necessary for transferring arrangements across scenes [39, 69]. As such, our task requires a holistic understanding of scene context, allowing for applications where spatial continuity and hierarchical understanding are critical. One example is in imitation learning for robotics and AR/VR [12, 13], where scene-to-scene task transfer can be more practical than generalizing control policies across environments.

Despite its practical benefits, the 3D scene analogy task poses unique challenges not addressed by traditional correspondence methods. First, a lack of dense ground-truth training data complicates learning, as contextual information varies widely across near-infinite scene configurations. Second, the task demands holistic reasoning about object relationships and surrounding open spaces at the point level, extending beyond conventional keypoint or scene graph matching methods, which often simplify objects as sparse keypoints or bounding boxes [60, 61, 37, 102, 15, 19]. Finally, robustness to appearance variation is crucial for managing cross-domain differences effectively.

As an effective solution to the 3D scene analogy problem, we introduce neural contextual scene maps. For a pair of 3D scenes, our method builds descriptor fields that capture detailed spatial relationships and finds matches by aligning the fields using a smooth map. Input to our method are sparsely sampled scene keypoints and their semantic information, resulting in a lightweight pipeline robust to input variations, noisy geometry, and appearance changes. Then the descriptor fields gather vicinity information to extract context-aware features. The fields are trained with contrastive learning, eliminating the need for densely labeled ground-truth data or inductive biases. Finally, our method estimates a smooth map aligning the descriptor fields through a coarse-to-fine procedure, which reduces the dependence on individual keypoints to reason about the overarching regional relations holistically.

Our approach effectively identifies accurate scene analogies for complex indoor scenes including noisy 3D scans, and is applicable to practical downstream tasks. Quantitative results show that our method outperforms baselines using vision foundation models [7, 67, 14, 19] or scene graphs [97, 75] on both real and synthetic 3D scenes, despite using a smaller feature dimension and training data. Additionally, our method also supports mapping *between* real and synthetic scenes indicating its robustness against input domain variations. We further demonstrate that our pipeline can be used for downstream tasks such as motion trajectory transfer and object placement, which can be extended to transfer long-term demo trajectories for robotics or create co-presence experiences for AR/VR applications.

To summarize, our main contributions are: i) introducing the 3D scene analogy task to find dense mappings between scene regions with common contexts, ii) developing neural contextual scene maps that combine spatial and semantic contexts of 3D keypoints to create smooth, detailed maps, and iii) demonstrating our method's generalizability across various inputs and applications.

II. Related Work

a) Instance and Group Correspondences: While the 3D scene analogy task is fairly new, there is extensive research on related problems in correspondence estimation, categorized by input settings and granularity. On the instance level, sparse matching methods in 2D (i.e., *semantic correspondence*) [62, 49, 19] and 3D [88, 68, 15] extract neural network features at keypoints to match between instances within the same semantic category. Similarly, dense matching methods in 2D (also known as *semantic flow*) [47, 46, 36, 48, 32, 38, 37, 65] exploit dense features and correlate them in the entire image space for matching. On the other hand, dense matching methods in 3D [64, 50, 63, 23] often start by finding sparse correspondences and optimizing smooth surface maps that pass through them. Our approach extends instance-level dense matching methods to finding *dense maps* over scene regions in 3D sharing similar contexts.

Unlike instance-level, most approaches in group-level correspondence target keypoint or object-wise matches. In the 2D case, multi-instance semantic correspondence [54, 81, 72, 99] aims to link sparse keypoints from an object instance in one image to multiple corresponding instances in another. For 3D, scene graph matching [75, 87, 97, 58] seeks correspondences between graphs representing 3D objects as nodes and their relationships as edges [2]. In contrast to these methods focusing on sparse matches, our work finds dense maps of contextually corresponding regions, accounting for both near-surface points and open spaces.

b) Neural Fields: Neural fields are spatio-temporal quantities that are parameterized fully or partially by a neural network [98]. Prominent applications of neural fields include photorealistic 3D reconstruction [59, 30, 11], 3D geometry extraction [91, 100], and SLAM [105, 89, 90]. While these studies primarily focus on visual fidelity and geometric accuracy, more recent works apply neural fields to semantic scene understanding [23, 35, 103] and robot motion planning [80, 92, 93, 79, 78]. Notably, studies on robot manipulation build fields using features from vision foundation models [7, 67, 14, 104] for establishing matches between observations during training and deployment. Our work aims to establish dense, context-aware correspondences that extend beyond visual/geometric fidelity or specific tasks such as manipulation. Further, while recent works in robotics [104, 79] consider transfer methods for *single* objects, our work enables transfer between *multiple* objects, encouraging future robotics research on multi-object demonstration transfer. Utilizing an efficient neural field based on sparse 3D keypoints, we achieve precise matches for both near-surface and open-space regions, which is difficult to attain from existing works.

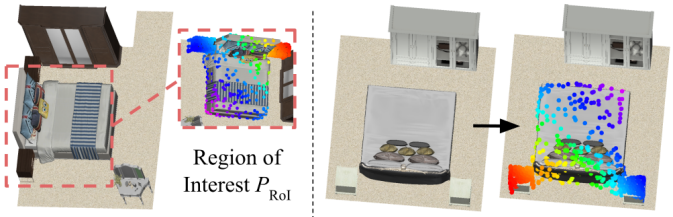


Fig. 2: Overview of our approach. Given a region of interest from object groups in the target scene, our method finds a smooth map to the corresponding region in the reference scene.

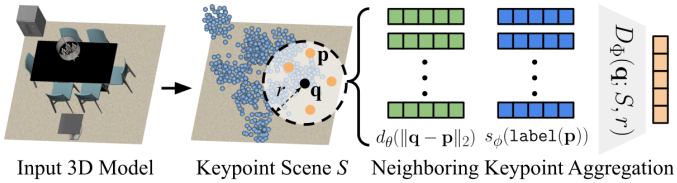


Fig. 3: Overview of context descriptor fields. Using sparsely sampled keypoints as the scene representation, for an arbitrary query point $\mathbf{q}$, the field gathers points within a radius r and computes a Transformer embedding based on the distance embedding $d_\theta(\|\mathbf{q}-\mathbf{p}\|_2)$ and semantic embedding $s_\phi(\texttt{label}(\mathbf{p}))$.

III. Method: Neural Contextual Scene Maps

Given a pair of scenes, our method finds a mapping from a region of interest in one scene to the corresponding region with similar scene contexts in the other scene (Figure 2). From a sparse set of points sampled in 3D scene models (Section III-A), our method first builds context descriptor fields that summarize the nearby geometry and semantic information for arbitrary query points (Section III-B). Based on descriptor fields, our method finds the dense map in a coarse-to-fine manner, by first extracting an affine map followed by local displacement maps (Section III-C).

A. Input Setup

a) Region of Interest (RoI) Representation: Let S_{tgt} denote the *target* scene, where the region of interest (RoI) is chosen, and S_{ref} the *reference* scene, to which the target scene region is mapped. As shown in Figure 2, we represent the RoI as a set of points $P_{\text{RoI}} \subset \mathbb{R}^3$, sampled from the surface of the object group we aim to match.

We then define the **neural contextual scene map** as a mapping $F(\cdot) : \text{conv}(S_{\text{tgt}}) \to \text{conv}(S_{\text{ref}})$, where $\text{conv}(S) \subset \mathbb{R}^3$ denotes the convex hull enclosing scene S. The scene map transforms points P_{RoI} to corresponding points in $\text{conv}(S_{\text{ref}})$ sharing similar scene contexts. Note, while we use specific points for feature encoding and loss calculation, the final output is a *dense map* across spatial regions, allowing us to find correspondences for any arbitrary point within the region. As an illustrative sample, Figure 1 shows our method mapping between a sofa-and-table group in the target scene to a similar object group in the reference scene.

b) Scene Representation: As shown in Figure 3, our method operates on a lightweight representation of scenes, using sparsely sampled keypoints from the original dense 3D model for efficiency. Formally, each scene is represented as a tuple $S = (\mathcal{O}, C)$ with an object set $\mathcal{O}$ and scene corner points $C \subset \mathbb{R}^3$. The object set $\mathcal{O} = \{(P_i, l_i)\}$ consists of points and semantic labels for each object in the scene where $P_i \subset \mathbb{R}^3$ denotes the point coordinates of the i^{th} object and $l_i \in \{1, \ldots, L\}$ denotes its semantic label among L classes. Scene corner points are either obtained from floorplan data if available [25, 26] or from points on convex hulls enclosing the scenes [18].

B. Context Descriptor Fields

Using sparse input representations, we design descriptor fields as lightweight scene representations that summarize scene context for arbitrary locations by aggregating nearby semantic and geometric information. For a scene S and a query point $\mathbf{q} \in \text{conv}(S)$, the context descriptor field $D_\Phi(\cdot) : \mathbb{R}^3 \to \mathbb{R}^d$ outputs a d-dimensional feature vector.

As shown in Figure 3, we implement the descriptor field using a Transformer encoder [85, 92]. The encoder first aggregates points in S that lie within distance r from $\mathbf{q}$, which we denote as $\mathcal{B}(\mathbf{q}; S, r)$. For each point $\mathbf{p} \in \mathcal{B}(\mathbf{q}; S, r)$, we concatenate a learned distance embedding $d_\theta(\|\mathbf{q}-\mathbf{p}\|_2)$ and a semantic embedding $s_\phi(\texttt{label}(\mathbf{p}))$ as Transformer input tokens. The descriptor field is defined as follows:

$$D_\Phi(\mathbf{q}; S, r) = \texttt{Transformer}(\{\texttt{Token}(\mathbf{p})\}_{\mathbf{p}\in\mathcal{B}(\mathbf{q};S,r)}), \quad (1)$$

where $\texttt{Token}(\mathbf{p}) = \texttt{Concat}(d_\theta(\|\mathbf{q}-\mathbf{p}\|_2), s_\phi(\texttt{label}(\mathbf{p})))$. To obtain the feature vector summarizing the input tokens, we append a learnable [CLS] token to the input token sequence in Equation 1 and use its output embedding as the final field vector [16, 17].

Descriptor fields holistically aggregate semantic and geometric information, enabling reasoning about fine-grained contextual correspondences. As an illustrative sample, Figure 4 shows the trained field distances between query points selected at open spaces in the target scene against uniformly sampled points in the reference scene. Notice sharp peaks are found only near chair arms next to the table corner (and not all chair arms), which indicates that descriptor fields can reason about detailed scene contexts.

1) Training Descriptor Fields: To train descriptor fields, we employ contrastive learning [10, 8, 9, 84, 96] on procedurally generated positive and negative scene pairs. Contrastive learning operates by maximizing the similarity of representations for positive data pairs with common attributes while minimizing similarity for negative pairs with dissimilar attributes. Since contrastive learning only requires positive and negative data pairs [8, 9, 96], our method can learn effective context-aware representations for descriptor fields without densely labeled training data, or hand-tuned inductive biases.

a) Dataset Generation: As shown in Figure 5, we propose an automated procedure to generate positive and negative scene pairs. Our pipeline assumes a *source* dataset consisting of 3D scenes $\mathcal{D}_{\text{src}}=\{S_i\}$ with known object poses. Among

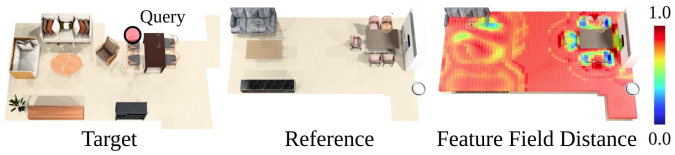


Fig. 4: Visualization of feature distances in open spaces. For a query point (red) in the target scene lying on the chair arm, we show the descriptor field distances against densely sampled points in the reference scene. Field values are only similar for chair arms near table corners, indicating that descriptor fields can reason about fine-grained contextual correspondences.

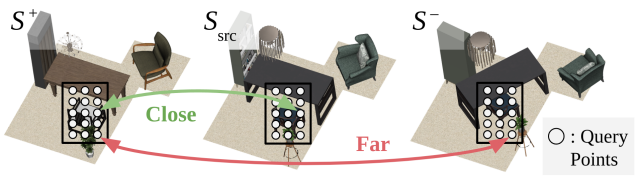


Fig. 5: Context descriptor field training overview. We replace each object in scene S_{src} with one with the same semantic label to create the positive scene S^+, and apply pose noise to obtain S^-. Contrastive learning is then applied to descriptor fields computed from points sampled within the object's bounding box.

the many possible definitions for a "correct" correspondence (e.g., appearance [76], style [56], or semantics [60]), we target finding point matches that share common nearby object semantics and local geometry, inspired from works in semantic correspondence [60, 61]. Based on this notion, the positive pairs (S_i, S_i^+) are generated by swapping objects in each scene S_i with randomly selected objects sharing the same semantic label from other scenes $\mathcal{D}_{\text{src}} \setminus S_i$. Here, the objects for replacement are sampled from the top-K (=100) list of objects having the most similar aspect ratios. Next, the negative pairs (S_i, S_i^-) are generated by adding noise perturbations to the object poses, similar to LEGO-Net [95]. Note that we constrain the pose noise to planar translation and z-axis rotation to prevent floating objects or ground penetrations. The resulting triplet dataset $\mathcal{D}_{\text{triplet}}=\{(S_i, S_i^+, S_i^-)\}$ is used for training.

b) Contrastive Learning: We extract query points from the generated scene triplets (S_i, S_i^+, S_i^-) for contrastive learning. Specifically, for each object in the source scene $o \in S_i$ and its corresponding object $o^+ \in S_i^+$, we sample an equal number of query points Q, Q^+ within the objects' oriented bounding box. Since objects o and o^+ share the same pose, we can associate each positive pair query point $\mathbf{q}^+$ with its corresponding source query point $\mathbf{q} \in Q$, as shown in Figure 5. Setting the negative query points as identical locations to the positive query points, the contrastive learning objective is defined as an InfoNCE loss [84, 8, 96] namely,

$$\mathcal{L} = \sum_{\mathbf{q},\mathbf{q}^+} -\log \frac{\exp(D_\Phi(\mathbf{q}; S, r)^T D_\Phi(\mathbf{q}^+; S^+, r)/\tau)}{\sum_{\tilde{S}\in\mathcal{S}} \exp(D_\Phi(\mathbf{q}; S, r)^T D_\Phi(\mathbf{q}^+; \tilde{S}, r)/\tau)}, \quad (2)$$

where $\mathcal{S} = \{S^+, S^-\}$ and τ is a temperature parameter set to 0.2 in all our experiments. Our training objective enforces the descriptor field to output similar embeddings for points lying on positive scene pairs and dissimilar embeddings for those on negative scene pairs. The trained fields are then used to estimate scene maps in the next section.

C. Contextual Scene Map Estimation

We now create a smooth map aligning the descriptor fields between two scenes. The design intentionally respects spatial vacancies and fine details near keypoints while reducing reliance on individual descriptors for enhanced robustness. Here, we employ a coarse-to-fine procedure to calculate the contextual map, as shown in Figure 6. Since there are many possible scene arrangements, the target and reference scenes

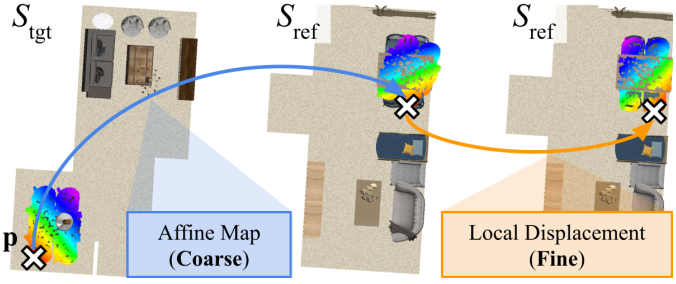


Fig. 6: Scene map estimation process overview. Our method first estimates affine maps to account for large transformations, and finds local displacements for detailed alignment.

may contain a different number of objects with shape variations. The coarse initialization with the smooth mapping can effectively ignore minor deviations between the two scenes and focus on deducing a holistic map. Specifically, we decompose the contextual scene map into an affine map and local displacements,

$$F(\mathbf{x}) := \mathbf{A}\mathbf{x} + \mathbf{b} + d_w(\mathbf{x}; P_{\text{RoI}}), \quad (3)$$

where $\mathbf{A} \in \mathbb{R}^{3\times3}$, $\mathbf{b} \in \mathbb{R}^3$ are the affine map parameters. We express the local displacement map as a linear combination of radial basis functions [33, 24],

$$d_w(\mathbf{x}; P_{\text{RoI}}) = \sum_k w_k \varphi(\|\mathbf{x} - \mathbf{p}_k\|), \quad (4)$$

where the control points are set as points on the RoI $\mathbf{p}_k \in P_{\text{RoI}}$ described in Section III-A. The basis function is set as the thin plate spline $\varphi(r) := r^2 \log(r)$. Intuitively, the affine map accounts for large, global transformations, and the local displacement map provides a fine-grained alignment for regions with similar contexts.

a) Affine Map Estimation: We extract a pool of affine maps by combinatorially associating object pairs in scenes S_{tgt} and S_{ref}, and optimize the initial maps from descriptor field alignment. Due to the low-dimensional structure of affine maps and the sparse keypoint representation, we can quickly select maps for further optimization. For each object pair $(o_{\text{tgt}}, o_{\text{ref}})$ with centroids $(\mathbf{c}_{\text{tgt}}, \mathbf{c}_{\text{ref}})$, we create a set of affine maps by associating object centroid displacements $\mathbf{c}_{\text{tgt}} - \mathbf{c}_{\text{ref}}$ with N_{ortho} uniformly sampled rotations and reflections in $SO(2)$. From the resulting $|\mathcal{O}_{\text{tgt}}| \times |\mathcal{O}_{\text{ref}}| \times N_{\text{ortho}}$ maps, we calculate the

following cost function for each affine map $(\mathbf{A}, \mathbf{b})$,

$$\mathcal{C}_{\text{coarse}} = \sum_{\mathbf{p} \in P_{\text{RoI}}} \| D_{\Phi}(\mathbf{p}; S_{\text{tgt}}) - D_{\Phi}(\mathbf{A}\mathbf{p} + \mathbf{b}; S_{\text{ref}}) \|, \quad (5)$$

where the descriptor fields are compared for points lying on the RoI P_{RoI}. Note we have omitted the radius input r for brevity. As the next step, we select K_{coarse} affine maps with the smallest cost values and perform a simple outlier object filtering procedure to remove objects in the RoI that are not matchable to the reference scene. After outlier removal, we optimize each affine map by minimizing the cost in Equation 5 with gradient descent [42, 51].

b) Local Displacement Map Estimation: We finally refine each affine map $(\mathbf{A}_{\text{opt}}, \mathbf{b}_{\textbf{opt}})$ selected from the previous step by further aligning fields with local displacements. Specifically, we minimize the following cost function

$$\mathcal{C}_{\text{fine}} = \sum_{\mathbf{p} \in P_{\text{RoI}}} \| D_{\Phi}(\mathbf{p}; S_{\text{tgt}}) - D_{\Phi}(\mathbf{A}_{\text{opt}}\mathbf{p} + \mathbf{b}_{\text{opt}} + \delta; S_{\text{ref}}) \| \quad (6)$$

where $\delta = d_w(\mathbf{p}; P_{\text{RoI}})$ is the local displacement defined in Equation 4. Similar to affine map estimation, we optimize the basis function weights w_k by minimizing Equation 6 with gradient descent. Finally, our method outputs the mapping with the smallest cost if the cost value is below a designated threshold ρ_{valid}, or otherwise labels the RoI to be *unmappable* to objects in the reference scene.

IV. Experiments

We evaluate our method for estimating 3D scene analogies on a wide range of 3D scenes (Section IV-A) and examine applicability in downstream tasks (Section IV-B).

a) Baselines: As finding 3D scene analogies is a new task, we compare our method against several contrived baselines, which are adaptations of recent 3D scene understanding pipelines [94, 82, 7, 67, 93, 75, 19]. First, the *scene graph matching* baseline constructs 3D scene graphs [2] and matches them via graph matching [87] followed by affine map estimation from object centroids. Next, the *multi-view semantic correspondence* baseline estimates 2D semantic correspondences between image rendering pairs of the input scenes using DINOv2 features [19, 67, 14], and lifts the 2D matches to 3D via back-projection.

The *visual feature field* and *3D point feature field* both generate 3D feature fields similar to our method and apply the map estimation from Section III-C. The visual feature field uses back-projected DINOv2 [67] features from scene renderings [93], while the 3D feature field extracts Vector Neuron [15] features for 3D keypoints and interpolates them for arbitrary queries [94].

b) Datasets: We evaluate two diverse indoor scene datasets: synthetic 3D scenes from 3D-FRONT [25] and real 3D scans from ARKitScenes [4] that include object semantic, instance, and pose labels suitable for training and evaluation. Context descriptor fields are trained separately on each dataset. We generate 10,000 training triplets for 3D-FRONT [25] using the procedure in Section III-B1 and 4,498 triplets for ARKitScenes [4] following the standard train/test split. In the absence of densely annotated ground-truth, we prepare two types of evaluation data to assess 3D scene analogies.

Metric	PCP		Bijectivity PCP		Chamfer Acc.	
Threshold	0.25	0.50	0.25	0.50	0.15	0.20
Scene Graph Matching	0.26	0.42	0.29	0.47	0.32	0.48
Multi-view Semantic Corresp.	0.10	0.20	0.14	0.21	0.62	0.86
Visual Feature Field	0.50	0.66	0.52	0.61	0.81	0.86
3D Point Feature Field	0.56	0.71	0.60	0.68	0.86	0.89
Ours	**0.76**	**0.90**	**0.92**	**0.94**	**0.97**	**0.99**

(a) Procedurally Generated Scene Pairs

Metric	Bijectivity PCP		Chamfer Acc.	
Threshold	0.25	0.50	0.15	0.20
Scene Graph Matching	0.22	0.36	0.27	0.40
Multi-view Semantic Corresp.	0.03	0.06	0.21	0.45
Visual Feature Field	0.56	0.58	0.69	0.75
3D Point Feature Field	0.53	0.56	0.64	0.69
Ours	**0.70**	**0.73**	**0.71**	**0.76**

(b) Manually Collected Scene Pairs

TABLE I: 3D scene analogy comparison in 3D-FRONT [25].

- **Procedurally generated scene pairs:** For each scene, we randomly select object groups and procedurally create a new scene containing them. Since object poses are known for the generated group matches, we apply the Hungarian algorithm [53] to obtain pseudo ground-truth maps.
- **Manually collected scene pairs:** We collect scene pairs with co-present object groups, along with pairs lacking common object groups to check whether any false positive 3D scene analogies are found.

c) Implementation Details: On both datasets, we extract object keypoints from the dense 3D model using farthest point sampling [20]. Scene corner points are obtained from the floorplan corners for 3D-FRONT [25, 26], and from convex hull points for ARKitScenes [4]. For descriptor fields, we set r=0.75 and d=256. During scene map estimation, we set N_{ortho}=16, K_{coarse}=5, ρ_{valid}=1.5, and optimize scene maps using Adam [51] with step size 10^{-3}.

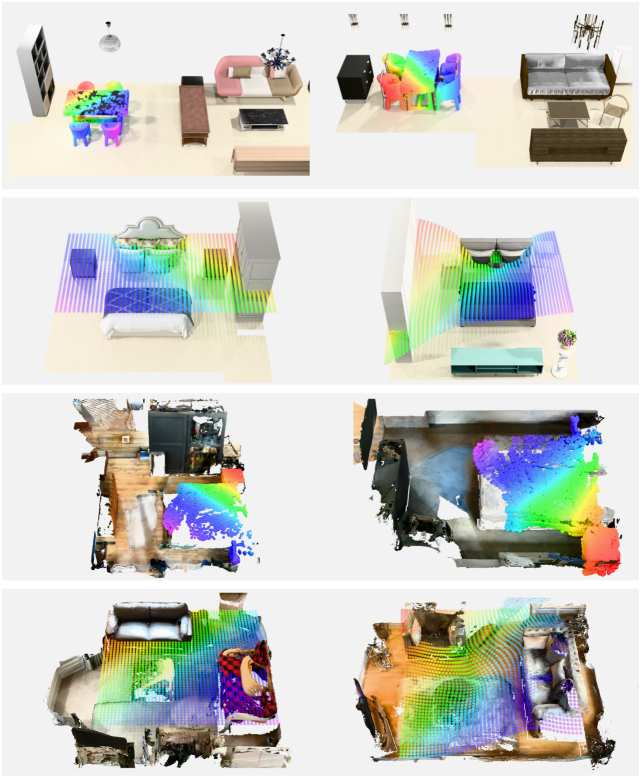
A. Performance Analysis

a) Metrics: We use three metrics for quantitative evaluation:

- **Percentage of Correct Points (PCP) [60, 61, 37, 102]:** This metric is used for procedurally generated scene pairs with pseudo ground-truth annotations. For points on the region of interest, the metric is defined as follows, $\text{PCP}(P_{\text{RoI}}) = 1/|P_{\text{RoI}}| \sum_{\mathbf{p}_{\text{RoI}} \in P} \mathbb{1}[\|F(\mathbf{p}) - \mathbf{p}_{\text{gt}}\| \leq \alpha]$, where α is a threshold parameter.
- **Bijectivity PCP [64, 63]:** After computing an inverse scene map $F^{-1}(\cdot) : S_{\text{ref}} \rightarrow S_{\text{tgt}}$ taking $F(P_{\text{RoI}})$ as input, this metric is defined as follows: $\text{Bi-PCP}(P_{\text{RoI}}) = 1/|P_{\text{RoI}}| \sum_{\mathbf{p} \in P_{\text{RoI}}} \mathbb{1}[\|F^{-1} \circ F(\mathbf{p}) - \mathbf{p}\| \leq \alpha]$.
- **Chamfer Accuracy:** The metric is defined as the percentage of predictions where i) the Chamfer distance [3, 22] between mapped points $F(P_{\text{RoI}})$ and sampled points in S_{ref} is below a threshold, or ii) no mappings are output for scene pairs with no common object groups.

The PCP and Bi-PCP metrics measure point-level accuracy of the estimated maps, while Chamfer accuracy evaluates group-level accuracy and penalizes false positive maps.

1) Scene Map Evaluation:



Target Reference

Fig. 7: Visualizations of estimated 3D scene analogies in 3D-FRONT and ARKitScenes. We show mapping results both for near-surface and open-space points.

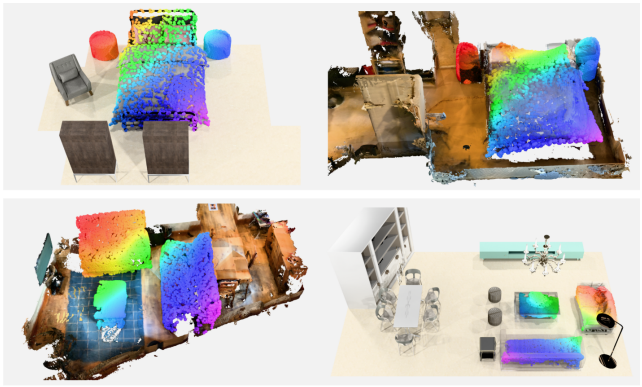
Metric	PCP		Bijectivity PCP		Chamfer Acc.	
Threshold	0.25	0.50	0.25	0.50	0.15	0.20
Scene Graph Matching	0.39	0.57	0.43	0.62	0.57	0.72
Multi-view Semantic Corresp.	0.10	0.21	0.10	0.18	0.59	0.78
Visual Feature Field	0.55	0.74	0.58	0.71	0.91	0.88
3D Point Feature Field	0.65	0.81	0.70	0.77	0.88	0.92
Ours	**0.75**	**0.90**	**0.90**	**0.94**	**0.96**	**0.99**

(a) Procedurally Generated Scene Pairs

Metric	Bijectivity PCP		Chamfer Acc.	
Threshold	0.25	0.50	0.15	0.20
Scene Graph Matching	0.25	0.37	0.33	0.45
Multi-view Semantic Corresp.	0.06	0.12	0.31	0.50
Visual Feature Field	0.26	0.29	0.40	0.42
3D Point Feature Field	0.41	0.49	0.51	0.60
Ours	**0.51**	**0.62**	**0.59**	**0.69**

(b) Manually Collected Scene Pairs

TABLE II: 3D scene analogy comparison in ARKitScenes [4].



Target Reference

Fig. 8: Visualizations of Sim2Real and Real2Sim scene analogies estimated between 3D-FRONT and ARKitScenes.

a) 3D-FRONT: We first present a quantitative comparison of the scene analogies found from our method with those from baselines on 3D-FRONT [25] in Table I. Our method consistently outperforms the baselines across all metrics for both procedurally generated and manually collected pairs. A large performance gap exists compared to the scene graph matching baseline, as it treats objects as single nodes, lacking geometric granularity. A similar trend is observed with the multi-view semantic correspondence baseline. While recent semantic correspondence methods excel at single object matching [60, 61, 19], they struggle to account for spatial relationships among multiple objects. Further, the feature field baselines based on DINOv2 [67] and Vector Neurons [15] also exhibit lower performance compared to our method, despite using the same coarse-to-fine map estimation process. Our method's contrastive learning pipeline enables effective descriptor extraction for highly accurate mappings, as shown in Figure 7.

b) ARKitScenes: We conduct further assessments on ARKitScenes [4], which, unlike 3D-FRONT, contains 3D scene meshes from real-world RGB-D camera measurements with noisy geometry and object layouts. As shown in Table II, our method outperforms baselines on most metrics, similar to 3D-FRONT [25], and generates accurate mappings as shown in Figure 7. This indicates that our descriptor fields and coarse-to-fine mapping scheme robustly handle the noisy inputs from real 3D scans. Nevertheless, all metrics show a consistent performance drop compared to the 3D-FRONT [25] results in Table I for manually collected scene pairs. We attribute this drop to largely incomplete geometry in several manually split scenes, which could be solved by modifying the cost functions in Equation 5, 6 to account for such outliers. While such a level of incompleteness is uncommon in real-world applications, addressing this issue is left for future work.

c) Sim2Real and Real2Sim Map Estimation: We investigate if our method can find analogies between synthetic 3D models in 3D-FRONT [25] and real scans in ARKitScenes [4]. Such capability is valuable for robotics and AR/VR applications: transferring pre-trained robot policies from virtual simulators to the real world [13], or enabling immersive telepresence by mapping real-world objects to their virtual counterparts [77]. Figure 8 shows estimated scene analogies for both sim-to-real and real-to-sim scenarios, using descriptor fields trained on 3D-FRONT [25] in both cases. The coarse-to-fine process allows holistic scene mapping, avoiding over-focus on individual descriptors and achieving reliable mappings across different domains.

2) Ablation Study:

a) Compatibility with Vision and Language Foundation Models: We assess our method's compatibility with vision and language foundation model features [70, 67, 16, 71] by training variants of the context descriptor fields using CLIP [70] or sentence embedding [71] in place of the semantic

Metric	Bijectivity PCP		Chamfer Acc.	
Threshold	0.25	0.50	0.15	0.20
Ours w/ CLIP Emb. [70]	0.77	0.81	0.91	**0.97**
Ours w/ Sentence Emb. [71]	0.78	0.82	0.92	**0.97**
Ours w/o Local Displacement	0.83	0.89	0.77	0.85
Ours	**0.90**	**0.92**	**0.94**	0.96

TABLE III: Ablation study of neural contextual scene maps, averaged on manual and procedural scene pairs from 3D-FRONT [25].

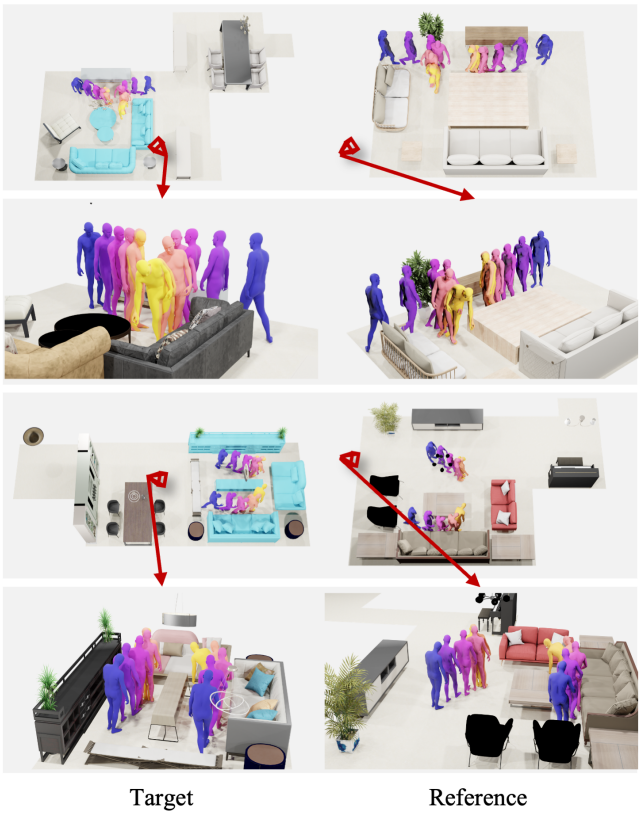


Fig. 9: Visualization of short trajectory transfer by directly mapping trajectory points. We shade the region of interest in blue.

embedding described in Section III-B. CLIP features are extracted from frontal view renders of each object in 3D-FRONT, while sentence embeddings are obtained by captioning each object renders with a vision-language model [1] and extracting text embeddings [71]. Table III shows scene map accuracy for manual and procedural scene pairs, with performance comparable to the original semantic embeddings and outperforming all the baselines. This shows that our method can effectively incorporate foundation model features without explicit semantic labels.

b) Local Displacement Maps: We finally ablate the coarse-to-fine mapping procedure by comparing our method to a variant that omits the local displacement estimation process. This results in suboptimal performance, as reported in Table III. Since mappings between scene regions with common contexts are often non-linear, relying solely on the affine map incurs inaccurate scene analogy detections.

B. Applications

a) Trajectory Transfer: Given a trajectory in an open space near the region of interest, we test if our method can

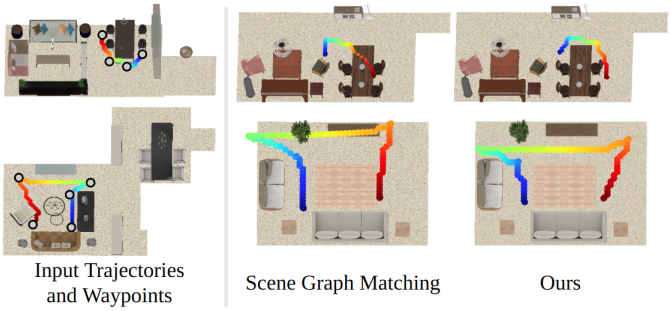
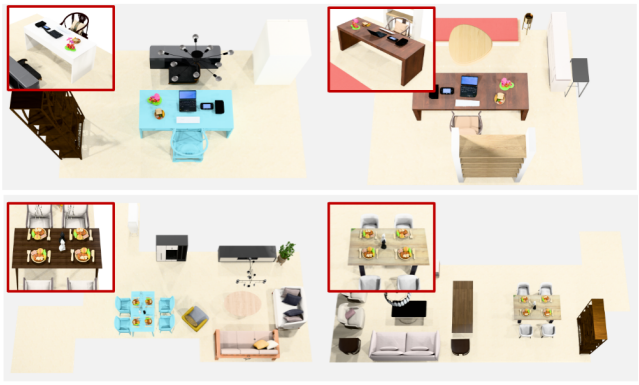


Fig. 10: Comparison of long trajectory transfer against scene graph matching. We estimate scene analogies to map waypoints, and apply traditional path planning [34] for interpolation.

transfer it to the reference scene's corresponding space. Such trajectory transfers aid in teleoperation [12], data augmentation / demonstration transfer for robot imitation learning [57, 104], or virtual co-presence [40] by mirroring the user's trajectory in a virtual environment. Our method can be applied flexibly depending on the length of the input trajectory. For short trajectories, we *directly* use the estimated map to transfer each trajectory point. Figure 9 visualizes a short trajectory transfer of virtual human agents moving through the target scene. We map the virtual human's bounding box corners at each timestamp and use the Umeyama algorithm [83] to find a rigid transformation to the reference scene. This result maintains consistent spatial relations with the surrounding objects: for example in Figure 9 the virtual human walking between a sofa and a table is accurately transferred. For long trajectories, directly using the estimated maps may cause collisions. We integrate our method with classical path planning [34, 55] by transferring sparse *waypoints* and finding collision-free paths using the A* algorithm [34] on the transferred waypoints. Figure 10 shows a comparison against scene graph matching. We interpolate object surface matches from the baseline to open space [86, 6] to find waypoint transfers, and directly apply the map when the A* algorithm fails due to inaccurate waypoint transfer. Compared to our method, this process results in erroneous transfers with penetrations. By producing a smooth map over $\mathbb{R}^3$, our method can flexibly handle trajectory transfer in open spaces, which is difficult with existing pipelines [75, 21, 19] that lack fine-grained understanding of spatial relations and surrounding context.

b) Object Placement Transfer: In contrast to trajectory and waypoint transfer, which focuses on *open space*, object placement transfer involves mapping small objects placed on a region of interest *surfaces* to the target scene. We first estimate scene maps from the region of interest and transfer objects placed on its surface via the scene map. The task is useful in AR/VR scenarios where users in different physical locations collaborate in a shared virtual space, allowing tools and objects in each user's space to align within the common virtual environment [77, 40, 43, 41, 44, 45]. As shown in Figure 11, a desk with small items can be accurately mapped from the target to the reference space. Our method successfully transfers objects



Target Reference

Fig. 11: Visualization of object placement transfer. We shade the region of interest used for scene analogy estimation in blue.

to coherent matching locations, demonstrating flexibility in handling both near-surface and open-space transfers.

V. Conclusion

We introduce *3D scene analogies*, which are dense maps between scenes with similar contexts, and propose neural contextual scene maps to find smooth, coherent mappings. Our method uses contextual descriptor fields and an effective coarse-to-fine estimation that holistically aligns the fields. Experiments demonstrate robustness across real-world scans and sim-to-real scenarios, with applications in trajectory and object placement transfer. We hope our work inspires future research in 3D scene context understanding.

a) Limitations: We acknowledge several limitations that invite future work. Currently, our method outputs a single mapping, whereas generating multiple plausible mappings could better account for symmetries and multi-modal correspondences. Additionally, while our method handles a wide range of spatial variations, it may struggle when object positions swap, as such changes disrupt the initial affine mapping. Future work could explore more flexible alignment strategies to address these cases. Finally, our evaluation is based on semantic and local geometric similarity, but different tasks may require alternative notions of "correctness." Expanding the evaluation framework to incorporate task-specific criteria could provide deeper insights.

References

[1] Moondream: A tiny vision model. https://moondream.ai/, 2024. Accessed: 2024-11-07.

[2] Iro Armeni, Zhi-Yang He, JunYoung Gwak, Amir R. Zamir, Martin Fischer, Jitendra Malik, and Silvio Savarese. 3d scene graph: A structure for unified semantics, 3d space, and camera. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[3] V. Athitsos and S. Sclaroff. Estimating 3d hand pose from a cluttered image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, volume 2, 2003. doi: 10.1109/CVPR.2003.1211500.

[4] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, and Elad Shulman. ARKitscenes - a diverse real-world dataset for 3d indoor scene understanding using mobile RGB-d data. In *Proceedings of the Conference on Neural Information Processing Systems Datasets and Benchmarks Track (NeurIPS)*, 2021. URL https://openreview.net/forum?id=tjZjv_qh_CE.

[5] Irving Biederman. Perceiving real-world scenes. *Science*, 177(4043):77–80, 1972. doi: 10.1126/science.177.4043.77. URL https://www.science.org/doi/abs/10.1126/science.177.4043.77.

[6] F.L. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 11(6):567–585, 1989. doi: 10.1109/34.24792.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[8] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

[9] Ting Chen, Simon Kornblith, Kevin Swersky, Mohammad Norouzi, and Geoffrey E Hinton. Big self-supervised models are strong semi-supervised learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 33, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/fcbc95ccdd551da181207c0c1400c655-Paper.pdf.

[10] Xinlei Chen and Kaiming He. Exploring simple siamese representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. URL https://api.semanticscholar.org/CorpusID:227118869.

[11] Zhiqin Chen, Thomas Funkhouser, Peter Hedman, and Andrea Tagliasacchi. Mobilenerf: Exploiting the polygon rasterization pipeline for efficient neural field rendering on mobile architectures. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[12] Xuxin Cheng, Jialong Li, Shiqi Yang, Ge Yang, and Xiaolong Wang. Open-television: Teleoperation with immersive active visual feedback. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2024.

[13] Tianyuan Dai, Josiah Wong, Yunfan Jiang, Chen Wang, Cem Gokmen, Ruohan Zhang, Jiajun Wu, and Li Fei-

Fei. Automated creation of digital cousins for robust policy learning. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2024.

[14] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[15] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J Guibas. Vector neurons: A general framework for so (3)-equivariant networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, June 2019. doi: 10.18653/v1/N19-1423. URL https://aclanthology.org/N19-1423.

[17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

[18] William F. Eddy. A new convex hull algorithm for planar sets. *ACM Transactions on Mathematical Software*, 3(4):398–403, December 1977. ISSN 0098-3500. doi: 10.1145/355759.355766. URL https://doi.org/10.1145/355759.355766.

[19] Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[20] Yuval Eldar, Michael Lindenbaum, Moshe Porat, and Yehoshua Zeevi. The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing*, 6:1305–15, 02 1997. doi: 10.1109/83.623193.

[21] Cathrin Elich, Iro Armeni, Martin R Oswald, Marc Pollefeys, and Joerg Stueckler. Learning-based relational object matching across views. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023. doi: 10.1109/ICRA48891.2023.10161393. URL https://doi.org/10.1109/ICRA48891.2023.10161393.

[22] Haoqiang Fan, Hao Su, and Leonidas Guibas. A Point Set Generation Network for 3D Object Reconstruction from a Single Image . In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017. doi: 10.1109/CVPR.2017.264. URL https://doi.ieeecomputersociety.org/10.1109/CVPR.2017.264.

[23] Michael Fischer, Zhengqin Li, Thu Nguyen-Phuoc, Aljaz Bozic, Zhao Dong, Carl Marshall, and Tobias Ritschel. Nerf analogies: Example-based visual attribute transfer for nerfs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[24] Richard Franke. Scattered data interpolation: tests of some methods. *Mathematics of Computation*, 38: 181–200, 1982. URL https://api.semanticscholar.org/CorpusID:8290519.

[25] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 3d-front: 3d furnished rooms with layouts and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[26] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. *International Journal of Computer Vision (IJCV)*, pages 1–25, 2021.

[27] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7 (2):155–170, 1983. ISSN 0364-0213. doi: https://doi.org/10.1016/S0364-0213(83)80009-3. URL https://www.sciencedirect.com/science/article/pii/S0364021383800093.

[28] Dedre Gentner and Virginia Gunn. Structural alignment facilitates the noticing of differences. *Memory & Cognition*, 29:565–577, 2001. URL https://api.semanticscholar.org/CorpusID:1745309.

[29] Dedre Gentner and Christian Hoyos. Analogy and abstraction. *Topics in cognitive science*, 9 3:672–693, 2017. URL https://api.semanticscholar.org/CorpusID:9307708.

[30] Lily Goli, Daniel Rebain, Sara Sabour, Animesh Garg, and Andrea Tagliasacchi. nerf2nerf: Pairwise registration of neural radiance fields. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2023.

[31] Katharine F Guarino, Elizabeth M Wakefield, Robert G Morrison, and Lindsey E Richland. Why do children struggle on analogical reasoning tasks? considering the role of problem format by measuring visual attention. *Acta Psychologica*, 224:103505, 2022.

[32] Kamal Gupta, Varun Jampani, Carlos Esteves, Abhinav Shrivastava, Abhinav Makadia, Noah Snavely, and Abhishek Kar. Asic: Aligning sparse image collections. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

[33] Rolland L. Hardy. Multiquadric equations of topography and other irregular surfaces. *Journal of Geophysical Research. 76 (8): 1905–1915*, 1971.

[34] Peter E. Hart, Nils J. Nilsson, and Bertram Raphael. A

formal basis for the heuristic determination of minimum cost paths. *IEEE Transactions on Systems Science and Cybernetics*, 4(2):100–107, 1968. doi: 10.1109/TSSC.1968.300136.

[35] Sunghwan Hong, Jisu Nam, Seokju Cho, Susung Hong, Sangryul Jeon, Dongbo Min, and Seungryong Kim. Neural matching fields: Implicit representation of matching fields for visual correspondence. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 35, 2022.

[36] Sangryul Jeon, Seungryong Kim, Dongbo Min, and Kwanghoon Sohn. Parn: Pyramidal affine regression networks for dense semantic correspondence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[37] Sangryul Jeon, Dongbo Min, Seungryong Kim, Jihwan Choe, and Kwanghoon Sohn. Guided semantic flow. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020. ISBN 978-3-030-58603-4. doi: 10.1007/978-3-030-58604-1_38. URL https://doi.org/10.1007/978-3-030-58604-1_38.

[38] Yuanchen Ju, Kaizhe Hu, Guowei Zhang, Gu Zhang, Mingrun Jiang, and Huazhe Xu. Robo-abc: Affordance generalization beyond categories via semantic correspondence for robot manipulation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[39] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the International Conference on Computer Vision (ICCV)*, 2023.

[40] Mohammad Keshavarzi, Michael Zollhoefer, Allen Y. Yang, Patrick Peluse, and Luisa Caldas. Synthesizing novel spaces for remote telepresence experiences. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, 2022. doi: 10.1109/ISMAR-Adjunct57072.2022.00111.

[41] Mohammad Keshavarzi, Michael Zollhoefer, Allen Yuqing Yang, Patrick Peluse, and Luisa Caldas. Mutual scene synthesis for mixed reality telepresence. *ArXiv*, abs/2204.00161, 2022. URL https://api.semanticscholar.org/CorpusID:247922742.

[42] J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Annals of Mathematical Statistics*, 23(3):462–466, 09 1952. doi: 10.1214/aoms/1177729392. URL https://doi.org/10.1214/aoms/1177729392.

[43] Dooyoung Kim, Hyung-Il Kim, and Woontack Woo. Mutual space generation with relative translation gains in redirected walking for asymmetric remote collaboration. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, 2022. doi: 10.1109/ISMAR-Adjunct57072.2022.00140.

[44] Dooyoung Kim, Seonji Kim, Jae-eun Shin, Boram Yoon, Jinwook Kim, Jeongmi Lee, and Woontack Woo. The effects of spatial configuration on relative translation gain thresholds in redirected walking. *Virtual Reality*, 27(2):1233–1250, December 2022. ISSN 1359-4338. doi: 10.1007/s10055-022-00734-3. URL https://doi.org/10.1007/s10055-022-00734-3.

[45] Seonji Kim, Dooyoung Kim, Jae-Eun Shin, and Woontack Woo. Object cluster registration of dissimilar rooms using geometric spatial affordance graph to generate shared virtual spaces. In *Proceedings of the IEEE Conference Virtual Reality and 3D User Interfaces (VR)*, 2024. doi: 10.1109/VR58804.2024.00099.

[46] Seungryong Kim, Dongbo Min, Bumsub Ham, Sangryul Jeon, Stephen Lin, and Kwanghoon Sohn. Fcss: Fully convolutional self-similarity for dense semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[47] Seungryong Kim, Dongbo Min, Stephen Lin, and Kwanghoon Sohn. Dctm: Discrete-continuous transformation matching for semantic flow. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[48] Seungryong Kim, Stephen Lin, Sang Ryul Jeon, Dongbo Min, and Kwanghoon Sohn. Recurrent transformer networks for semantic correspondence. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, volume 31, 2018.

[49] Seungwook Kim, Juhong Min, and Minsu Cho. Transformatcher: Match-to-match attention for semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[50] Vladimir G Kim, Yaron Lipman, and Thomas Funkhouser. Blended intrinsic maps. *ACM Transactions on Graphics (TOG)*, 30(4):1–12, 2011.

[51] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun, editors, *Proceedings of the International Conference on Learning Representations (ICLR)*, 2015. URL http://arxiv.org/abs/1412.6980.

[52] Daniel C. Krawczyk. The cognition and neuroscience of relational reasoning. *Brain Research*, 1428:13–23, 2012. ISSN 0006-8993. doi: https://doi.org/10.1016/j.brainres.2010.11.080. URL https://www.sciencedirect.com/science/article/pii/S000689931002593X.

[53] Harold W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics (NRL)*, 52, 1955. URL https://api.semanticscholar.org/CorpusID:9426884.

[54] Shiyi Lan, Zhiding Yu, Christopher Choy, Subhashree Radhakrishnan, Guilin Liu, Yuke Zhu, Larry S Davis, and Anima Anandkumar. Discobox: Weakly supervised instance segmentation and semantic correspondence from box supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*,

2021.

[55] S.M. LaValle and J.J. Kuffner. Randomized kinodynamic planning. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, volume 1, pages 473–479, 1999. doi: 10.1109/ROBOT.1999.770022.

[56] Kunhao Liu, Fangneng Zhan, Yiwen Chen, Jiahui Zhang, Yingchen Yu, Abdulmotaleb El Saddik, Shijian Lu, and Eric Xing. Stylerf: Zero-shot 3d style transfer of neural radiance fields. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[57] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: A data generation system for scalable robot learning using human demonstrations. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2023.

[58] Yang Miao, Francis Engelmann, Olga Vysotska, Federico Tombari, Marc Pollefeys, and Dániel Béla Baráth. Scenegraphloc: Cross-modal coarse visual localization on 3d scene graphs. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2024.

[59] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 65(1): 99–106, 2021.

[60] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Hyperpixel flow: Semantic correspondence with multi-layer neural features. In *ICCV*, 2019.

[61] Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. Spair-71k: A large-scale benchmark for semantic correspondence. *ArXiv*, abs/1908.10543, 2019.

[62] Juhong Min, Seungwook Kim, and Minsu Cho. Convolutional hough matching networks for robust and efficient visual correspondence. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 45 (7):8159–8175, 2023.

[63] Luca Morreale, Noam Aigerman, Vladimir G Kim, and Niloy J Mitra. Neural surface maps. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[64] Luca Morreale, Noam Aigerman, Vladimir G Kim, and Niloy J Mitra. Neural semantic surface maps. *Computer Graphics Forum*, 43(2):e15005, 2024.

[65] Jisu Nam, Gyuseong Lee, Sunwoo Kim, Hyeonsu Kim, Hyoungwon Cho, Seyeon Kim, and Seungryong Kim. Diffusion model for dense matching. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2023.

[66] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527, 2007. ISSN 1364-6613. doi: https://doi.org/10.1016/j.tics.2007.09.009. URL https://www.sciencedirect.com/science/article/pii/S1364661307002550.

[67] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, Po-Yao Huang, Hu Xu, Vasu Sharma, Shang-Wen Li, Wojciech Galuba, Mike Rabbat, Mido Assran, Nicolas Ballas, Gabriel Synnaeve, Ishan Misra, Herve Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2023.

[68] Chunghyun Park, Seungwook Kim, Jaesik Park, and Minsu Cho. Learning so (3)-invariant semantic correspondence via local shape transform. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[69] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Openscene: 3d scene understanding with open vocabularies. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[70] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[71] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019. URL https://arxiv.org/abs/1908.10084.

[72] Zhou Ren, Hailin Jin, Zhe Lin, Chen Fang, and Alan Yuille. Multi-instance visual-semantic embedding. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2015.

[73] Lindsey E. Richland, Robert G. Morrison, and Keith J. Holyoak. Children's development of analogical reasoning: Insights from scene analogy problems. *Journal of Experimental Child Psychology*, 94(3):249–273, 2006. ISSN 0022-0965. doi: https://doi.org/10.1016/j.jecp.2006.02.002. URL https://www.sciencedirect.com/science/article/pii/S0022096506000245.

[74] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2022.

[75] Sayan Deb Sarkar, Ondrej Miksik, Marc Pollefeys, Daniel Barath, and Iro Armeni. Sgaligner : 3d scene alignment with scene graphs. *Proceedings of the IEEE/CVF International Conference on Computer Vi-*

sion (ICCV), 2023.

[76] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *Proceedings of the IEEE/CVR Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[77] HyunA Seo, Juheon Yi, Rajesh Balan, and Youngki Lee. Gradualreality: Enhancing physical object interaction in virtual reality via interaction state-aware blending. In *Proceedings of the Annual ACM Symposium on User Interface Software and Technology (UIST)*, 2024.

[78] Nur Muhammad Mahi Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. In *Proceedings of the Workshop on Language and Robotics at the Conference on Robot Learning (CoRL)*, 2022.

[79] William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2023. URL https://openreview.net/forum?id=Rb0nGIt_kh5.

[80] Anthony Simeonov, Yilun Du, Andrea Tagliasacchi, Joshua B Tenenbaum, Alberto Rodriguez, Pulkit Agrawal, and Vincent Sitzmann. Neural descriptor fields: Se (3)-equivariant object representations for manipulation. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*, 2022.

[81] Yixuan Sun, Yiwen Huang, Haijing Guo, Yuzhou Zhao, Runmin Wu, Yizhou Yu, Weifeng Ge, and Wenqiang Zhang. Misc210k: A large-scale dataset for multi-instance semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[82] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2023. URL https://openreview.net/forum?id=ypOiXjdfnU.

[83] S. Umeyama. Least-squares estimation of transformation parameters between two point patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 13(4):376–380, 1991. doi: 10.1109/34.88573.

[84] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. URL https://api.semanticscholar.org/CorpusID:49670925.

[85] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the Conference on Neural Information Processing Systems (NeurIPS)*, 2017. ISBN 9781510860964.

[86] Grace Wahba. Spline models for observational data. In *CBMS-NSF Regional Conference Series in Applied Mathematics*, 1990. URL https://api.semanticscholar.org/CorpusID:121858740.

[87] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[88] Hanyu Wang, Jianwei Guo, Dong-Ming Yan, Weize Quan, and Xiaopeng Zhang. Learning 3d keypoint descriptors for non-rigid shape matching. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018.

[89] Hengyi Wang, Jingwen Wang, and Lourdes Agapito. Co-slam: Joint coordinate and sparse parametric encodings for neural real-time slam. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[90] Jingwen Wang, Tymoteusz Bleja, and Lourdes Agapito. Go-surf: Neural feature grid optimization for fast, high-fidelity rgb-d surface reconstruction. In *Proceedings of the IEEE International Conference on 3D Vision (3DV)*, 2022.

[91] Peng Wang, Lingjie Liu, Yuan Liu, Christian Theobalt, Taku Komura, and Wenping Wang. Neus: Learning neural implicit surfaces by volume rendering for multi-view reconstruction. In *Proceedings of the Conference on Neural Information Processing Systems*, 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/e41e164f7485ec4a28741a2d0ea41c74-Paper.pdf.

[92] Qianxu Wang, Congyue Deng, Tyler Ga Wei Lum, Yuanpei Chen, Yaodong Yang, Jeannette Bohg, Yixin Zhu, and Leonidas Guibas. Neural attention field: Emerging point relevance in 3d scenes for one-shot dexterous grasping. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2024. URL https://arxiv.org/abs/2410.23039.

[93] Qianxu Wang, Haotong Zhang, Congyue Deng, Yang You, Hao Dong, Yixin Zhu, and Leonidas Guibas. SparseDFF: Sparse-view feature distillation for one-shot dexterous manipulation. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=HHWlwxDeRn.

[94] Yixuan Wang, Mingtong Zhang, Zhuoran Li, Tarik Kelestemur, Katherine Driggs-Campbell, Jiajun Wu, Li Fei-Fei, and Yunzhu Li. D^3fields: Dynamic 3d descriptor fields for zero-shot generalizable rearrangement. In *Proceedings of the Conference on Robot Learning (CoRL)*, 2024.

[95] Qiuhong Anna Wei, Sijie Ding, Jeong Joon Park, Rahul Sajnani, Adrien Poulenard, Srinath Sridhar, and Leonidas Guibas. Lego-net: Learning regular rearrangements of objects in rooms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern*

Recognition (CVPR), June 2023.

[96] Saining Xie, Jiatao Gu, Demi Guo, Charles R. Qi, Leonidas Guibas, and Or Litany. Pointcontrast: Unsupervised pre-training for 3d point cloud understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[97] Yaxu Xie, Alain Pagani, and Didier Stricker. SG-PGM: Partial Graph Matching Network with Semantic Geometric Fusion for 3D Scene Graph Alignment and its Downstream Tasks . In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.02683.

[98] Yiheng Xie, Towaki Takikawa, Shunsuke Saito, Or Litany, Shiqin Yan, Numair Khan, Federico Tombari, James Tompkin, Vincent Sitzmann, and Srinath Sridhar. Neural fields in visual computing and beyond. *Computer Graphics Forum*, 41:641–676, 2022. ISSN 1467-8659. doi: 10.1111/cgf.14505.

[99] Xiangyang Xue, Wei Zhang, Jie Zhang, Bin Wu, Jianping Fan, and Yao Lu. Correlative multi-label multi-instance image annotation. In *Proceedings of the International Conference on Computer Vision (ICCV)*. IEEE, 2011.

[100] Lior Yariv, Yoni Kasten, Dror Moran, Meirav Galun, Matan Atzmon, Basri Ronen, and Yaron Lipman. Multiview neural surface reconstruction by disentangling geometry and appearance. *Proceedings of the Conference on Neural Information Processing Systems*, 2020.

[101] Yang You, Yujing Lou, Ruoxi Shi, Qi Liu, Yu-Wing Tai, Lizhuang Ma, Weiming Wang, and Cewu Lu. PRIN/SPRIN: On Extracting Point-Wise Rotation Invariant Features . *IEEE Transactions on Pattern Analysis & Machine Intelligence (TPAMI)*, 44(12):9489–9502, December 2022. ISSN 1939-3539. doi: 10.1109/TPAMI.2021.3130590. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2021.3130590.

[102] Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024.

[103] Shuaifeng Zhi, Tristan Laidlow, Stefan Leutenegger, and Andrew J. Davison. In-place scene labelling and understanding with implicit scene representation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[104] Junzhe Zhu, Yuanchen Ju, Junyi Zhang, Muhan Wang, Zhecheng Yuan, Kaizhe Hu, and Huazhe Xu. Densematcher: Learning 3d semantic correspondence for category-level manipulation from a single demo. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2025. URL https://openreview.net/forum?id=8oFvUBvF1u.

[105] Zihan Zhu, Songyou Peng, Viktor Larsson, Weiwei Xu, Hujun Bao, Zhaopeng Cui, Martin R Oswald, and Marc Pollefeys. Nice-slam: Neural implicit scalable encoding for slam. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, 2022.