CrossWeave: Cross-Modal Scene and Instance Retrieval from Sparse 2D-3D Alignment

Aadith Warrier[1] **Gnana Prakash Punnavajhala**[1] **Siddharth Tourani**[2]
Muhammad Haris Khan[3] **Avinash Sharma**[4] **Madhava Krishna**[1]

[1]IIIT Hyderabad [2]University of Heidelberg [3]MBZUAI [4]IIT Bombay

crossweave3d.github.io github.com/crossweave3d/crossweave

Abstract

Cross-modal scene and instance retrieval aims to retrieve a corresponding 3D environment or object given query 2D images, serving as a foundational mechanism for spatial reasoning in robotics and AR/VR. Existing methods achieve this by learning alignment only at the level of whole scenes or objects, which provides no supervision for the fine-grained local structure within them. In this work, we show that supervision on a set of sparse 2D-3D correspondences, without scene- or instance-level contrastive objectives, is sufficient to train a single model for both scene- and instance-level retrieval. Specifically, we propose **CrossWeave**, a novel framework for cross-modal retrieval that learns sparse patch-point alignment, which alone outperforms existing methods. An optimal transport-based aggregator over the aligned features gives further gains by clustering structurally redundant regions, freeing descriptor capacity for semantically distinct local features. CrossWeave achieves state-of-the-art performance on both scene- and instance-level retrieval on ScanNet and 3RScan with a *single model*, whereas existing methods require separate models for each task. Our results suggest that whole-scene encoding is unnecessary for cross-modal retrieval and that sparse local correspondence supervision is a more effective and flexible alternative.

1 Introduction

Within autonomous navigation and mapping, processing continuous streams of image and point cloud data is a major computational bottleneck. Scene and instance retrieval provide a first-pass spatial reasoning step, enabling location- and object-specific downstream tasks without exhaustive search over large databases or 3D maps. Such coarse reasoning is often sufficient for identifying object presence, recognizing the current environment, or initializing fine-grained localization and loop closure in SLAM. Cross-modal retrieval is especially valuable when the map is built using a richer sensor modality than the query device, such as a camera-equipped robot localizing against a LiDAR-built map, or an AR/VR device anchoring content to a pre-scanned environment.

Formally, **Cross-Modal Scene Retrieval (CMSR)** is defined as retrieving, from a database of aligned 3D point cloud sub-maps, the sub-map corresponding to query RGB image(s). The related task of **Cross-Modal Instance Retrieval (CMIR)** is defined analogously: given RGB image(s) of an object, the system retrieves the corresponding 3D point cloud instance from a database of segmented object point clouds. CMSR remains an emerging problem, with few established methods for bridging the modality gap between 2D images and 3D maps. Existing approaches such as *CrossOver* [27] and *SceneGraphLoc* [19] rely on scene- or object-level supervision: CrossOver aligns modalities with separate scene- and instance-level contrastive objectives and requires a separate model for each task, while SceneGraphLoc matches image patches to the object nodes of a multimodal scene graph. In

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

both cases, descriptors are learned using contrastive objectives similar to those used in visual place recognition [1, 11]. However, both methods supervise alignment at the level of whole scenes or objects rather than individual 3D points.

To address this, we propose **CrossWeave**, a cross-modal retrieval framework that learns from sparse local scene regions during training instead of aligning full scenes. CrossWeave introduces an image-patch-to-3D-point contrastive objective that preserves local cross-modal correspondences. It further aligns corresponding image-point cloud pairs using modality-specific descriptors obtained by aggregating local features, without requiring full-scene alignment. Using this architecture, we show that a single model trained with these losses, without any scene- or instance-level supervision, supports both scene- and object-level retrieval on the 3RScan and ScanNet benchmarks. We further demonstrate that patch-point alignment is the primary driver of instance retrieval performance, and improves scene retrieval when combined with learned aggregation.

Contributions:

- We propose **CrossWeave**, a novel multi-modal framework that performs both CMSR and CMIR with a single model, using a dual-contrastive loss at both the local patch-point level and the image-point cloud level, but not a scene/instance-level contrastive loss. This allows improved retrieval when compared to methods that encode the whole 3D map either as global point cloud features or scene graphs.
- We show that adding local patch-point alignment consistently improves both CMSR and CMIR over training on aggregated descriptors alone, with the largest gains on fine-grained instance retrieval.
- We achieve state-of-the-art performance on 3RScan and ScanNet, outperforming competing baselines by 21 points on scene retrieval and 8 points on instance retrieval on ScanNet.

2 Related Work

2.1 Multi-modal Representation Learning

Multi-modal representation learning aims to align heterogeneous data sources within a shared embedding space. In this setting, contrastive learning has become the dominant paradigm. CLIP [25] demonstrated that large-scale contrastive training on paired image-text data yields rich representations capable of cross-modal understanding and strong zero-shot generalization to various downstream tasks. Building on this foundation, a series of works extended cross-modal alignment to the image-point cloud [40, 44] and text-image-point cloud modalities [35, 36, 15] with Uni3D [43] scaling model and dataset sizes used for training. Subsequently, ImageBind [9] performed alignment between six modalities by leveraging images as a hub and demonstrated emergent relationships between untrained pairs. Similarly, PointBind [10] incorporated the 3D modality and performed various downstream tasks by augmenting with an LLM. However, these methods remained restricted to object-level point clouds. CLIP^2 [38] extended multi-modal representation learning to scene-level point clouds, yet remained focused on semantic understanding rather than retrieval. Overall, prior work focuses on cross-modal alignment for semantic understanding, with comparatively less emphasis on retrieval-oriented representation design for 3D scenes.

2.2 Uni-modal and Cross-modal Retrieval

Retrieval methods aim to identify relevant database entries given a query, typically represented by a global descriptor. While early methods aggregated local features via hand-crafted clustering [13], later works introduced learned aggregation like NetVLAD [1] to preserve distributional structure. Uni-modal visual [1, 14] and LiDAR-based [30, 31, 37, 24, 42] place recognition are widely studied along these lines. In the cross-modal domain, image-text retrieval is well-established [3], heavily driven by contrastive joint-embedding alignment [25, 17, 28]. Extending beyond 2D, recent works bridge textual queries with 3D point clouds via part-to-word matching [29] and multi-modal consistencies [26, 8]. Concurrently, scene graphs are used to capture explicit structural relations for cross-modal text-to-image retrieval, evolving from basic query parsing [12] to advanced message passing and attribute alignment [6, 33]. Despite these advances, cross-modal scene and instance retrieval remains largely underexplored, particularly for indoor environments mapped between images and 3D point clouds.

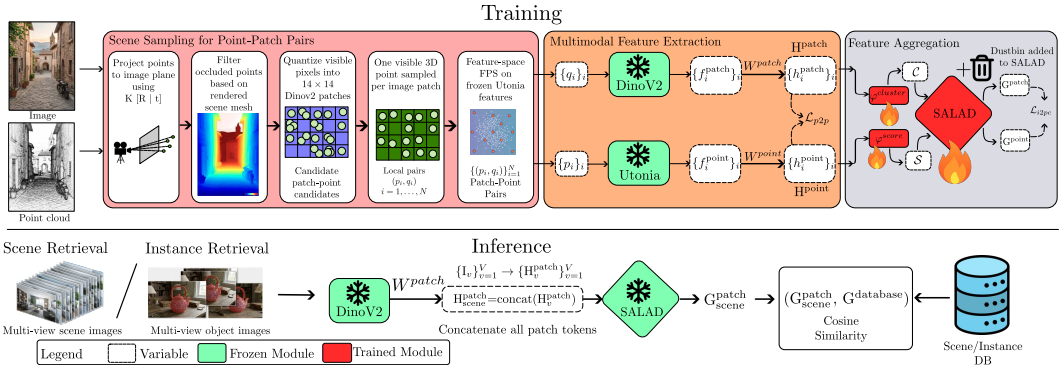


Figure 1: **Overview of CrossWeave.** CrossWeave learns cross-modal retrieval representations from sparse image patch-3D point correspondences rather than scene- or instance-level labels. **Scene Sampling:** Given calibrated RGB-D observations and a reconstructed 3D map, visible patch-point pairs are obtained by projecting 3D points into the image plane, filtering occluded points using rendered depth, quantizing valid pixels into DINOv2 patch tokens, and selecting a compact semantically diverse subset using feature-space farthest point sampling. **Feature Extraction:** Frozen encoders extract local patch and point features, which are projected by learnable projection matrices W^{patch} and W^{point} into a shared embedding space. **Feature Aggregation:** A shared SALAD aggregator then maps variable-sized sets of local features into fixed-dimensional descriptors. Training combines a local patch-point $\mathcal{L}_{p2p}$ contrastive loss with an aggregated image-point-cloud contrastive loss $\mathcal{L}_{i2pc}$, preserving fine-grained alignment while organizing local evidence into retrieval-ready descriptors. At inference, the same encoders and projection heads serve both tasks: SALAD descriptors are used for scene retrieval and mean-pooled local features for instance retrieval.

The most closely related work to ours is CrossOver [27], which trains for cross-modal alignment across text, images, floorplans, CAD models, and point clouds at both scene and instance levels, enabling retrieval across any pair of these modalities at multiple levels of granularity. We highlight two design aspects of prior work most relevant to our setting. First, instance- and scene-level retrieval rely on separate task-specific checkpoints, requiring distinct models for different granularities and limiting flexibility at inference time. Second, scene-level representations are constructed by averaging per-frame features (CLS and mean patch tokens) across frames for images and mean pooling point features for point clouds, which implicitly assigns equal weight to all local features, potentially losing fine-grained structural information critical for distinguishing visually similar scenes. In contrast, our method uses a unified model that learns fine-grained patch-to-point correspondences, which can be flexibly aggregated at inference time to support both instance- and scene-level retrieval. Furthermore, our aggregation assigns local features to shared cluster centers, capturing their distributional structure rather than averaging them away.

3 Methodology

The design philosophy behind CrossWeave is to both: effectively utilize the redundancy in the 3D scene structure and not lose fine-grained details due to the aggregation process over the local features to obtain modality-specific descriptors. Figure 1 provides an overview of CrossWeave.

3.1 Architecture and Training

In keeping with the convention in previous CMSR literature, we assume the 3D structure consists of individual RGB-D scans aligned into a single 3D map and not the individual scans themselves.

Scene Sampling for Patch-Point Pair Generation To exploit the redundancy present in both image and point-cloud observations, we avoid encoding all 3D points visible in an image and instead construct a compact set of informative patch-point correspondences. Given a point cloud P, we first project each 3D point into the image plane using the camera intrinsics K and camera pose $[R \mid t]$. Since projection alone does not guarantee that a point is actually visible from the camera viewpoint, we apply an explicit visibility constraint using the depth map rendered from the reconstructed

mesh. A projected point is retained only if its depth lies within a small tolerance ϵ of the observed depth at the corresponding pixel. This removes occluded points and ensures that the resulting correspondences reflect physically visible image-geometry associations. These correspondences are the only supervision used during training, and no scene or instance labels are required.

The remaining valid pixel-point pairs are then quantized into image patch indices using a patch size of 14, following the patchification scheme of DINOv2 [22]. For each image patch, we randomly sample a single visible 3D point, producing a one-to-one patch-point mapping. This step reduces the strong local redundancy caused by multiple 3D points projecting to the same visual token, while preserving the spatial correspondence required for patch-level cross-modal supervision.

Finally, we further downsample these patch-point pairs using Farthest Point Sampling (FPS) [7] in feature space rather than Euclidean space. Specifically, FPS is applied to the point features extracted by our frozen 3D encoder, Utonia [41], a recent point-cloud feature extractor based on the Point Transformer v3 architecture [34]. Unlike geometric FPS, which primarily encourages spatial coverage, feature-space FPS encourages semantic diversity by selecting points that are maximally separated in the learned embedding space. This prevents the sampled set from being dominated by large or frequently observed objects in the field of view, yielding a more balanced set of local correspondences for contrastive learning. The output of this scene sampling is a set of point-patch pairs $\{p_i \in \mathbb{R}^3, q_i \in \mathbb{R}^{14\times14}\}, i = 1 \to N$.

Feature Extraction The sampled point-patch pairs are passed on to pre-trained modality-specific encoders. For the image patch q_i, we use the DINOv2-L variant for generating per-patch features f_i^{patch}. For the 3D point p_i, we use Utonia to extract per-point features f_i^{point} from the point cloud. Subsequently, both features are ℓ_2 normalized and projected into a shared embedding space using learnable projection matrices W_{patch} and W_{point}. This results in per modality aligned features $H^{patch} = \{h_i^{\text{patch}}\}_{i=1}^N$ and $H^{point} = \{h_i^{\text{point}}\}_{i=1}^N$ per image.

Feature Aggregation via Optimal Transport For each image-point-cloud sample, we aggregate the N sampled patch-point correspondences into fixed-dimensional descriptors using a SALAD [11] aggregation layer. Here, N denotes the number of retained patch-point pairs after visibility filtering and sampling, while K denotes the number of latent aggregation clusters used by SALAD. The goal of this aggregation is to summarize a compact set of local cross-modal correspondences into a single descriptor, while capturing discriminative and recurring patterns in the cross-modal data. Importantly, aggregation is performed only over the N sampled correspondences from one image-point-cloud pair, rather than over all points in a full scene or features of an object instance.

To complete the mapping from patch-point level local features (H^{point}, H^{patch}) to image-point cloud level features, SALAD uses two MLPs shared across both modalities. The first MLP, φ^{cluster}, maps each local feature to a cluster feature, while the second MLP, φ^{score}, predicts its assignment scores to the K latent clusters:

$$C = \varphi^{\text{cluster}}(H) \in \mathbb{R}^{N\times d_{\text{cl}}}, \qquad S = \varphi^{\text{score}}(H) \in \mathbb{R}^{N\times K}.$$

Here, d_{cl} is the dimensionality of the transformed local feature used inside the aggregation layer. Thus, C_i, the i-th row of C, is the feature content contributed by the i-th local correspondence, while $S_{i,k}$ measures how strongly it should be assigned to the k-th latent cluster.

Following SALAD [11], the score matrix S is augmented with a learnable dustbin parameter ϕ and passed through the Sinkhorn algorithm [4] for three iterations. This produces a balanced soft-assignment matrix $\mathcal{P} \in \mathbb{R}^{K\times N}$, after removing the dustbin assignments. The dustbin allows low-confidence or uninformative local features to be absorbed rather than being forced into one of the K clusters, which improves robustness when some sampled correspondences are noisy or weakly discriminative. Each cluster descriptor V_k is then computed as a weighted sum of the transformed local features assigned to the k-th cluster. Thus, each cluster summarizes the subset of local evidence associated with the k-th latent concept. The vectors V_k are individually ℓ_2-normalized, concatenated, and then globally ℓ_2-normalized to obtain the final descriptor.

$$G = \text{norm}\Big(\text{concat}\big(\text{norm}(V_1), \ldots, \text{norm}(V_K)\big)\Big) \in \mathbb{R}^{Kd_{\text{cl}}}.$$

Applying this aggregation independently to H^{patch} and H^{point} yields the modality-specific descriptors G^{patch} and G^{point}, which are later used for cross-modal contrastive supervision.

Dual-Objective Contrastive Learning CrossWeave is trained with two complementary contrastive objectives, both based on InfoNCE [21, 20]. The first objective, the local patch-point loss $\mathcal{L}_{\text{p2p}}$, is computed over all $M = B \times N$ patch-point correspondences in a batch, where B is the batch size and N is the number of sampled correspondences per image. This loss directly aligns local visual and geometric features, encouraging each image patch to be close to its corresponding 3D point while treating other patch-point pairs in the batch as negatives. Since it is defined only at the correspondence level, it does not require scene-level or instance-level annotations.

The second objective, the image-point-cloud loss $\mathcal{L}_{\text{i2pc}}$ realized as the InfoNCE loss as well, is applied to the SALAD-aggregated descriptors of each image-point-cloud pair in the batch. This loss encourages the aggregated image descriptor and the corresponding aggregated point-cloud descriptor to occupy a shared retrieval space. Unlike full-scene supervision, this alignment is derived from descriptors built only from the sampled local correspondences, preserving the focus on sparse but discriminative cross-modal evidence.

The final training objective combines the local and aggregated contrastive losses, $\mathcal{L} = \mathcal{L}_{\text{p2p}} + \mathcal{L}_{\text{i2pc}}$. Together, these objectives allow CrossWeave to learn retrieval-oriented representations from sparse patch-point correspondences alone. The local loss preserves fine-grained cross-modal alignment, while the aggregated loss organizes these local cues into image-level and point-cloud-level descriptors.

3.2 Inference

Although CrossWeave is trained only on sparse local patch-point correspondences, it can be applied directly to both scene-level and instance-level retrieval at inference. This follows from the input-agnostic design of the model. The projection layers are applied independently to each local image or point feature, and the SALAD aggregator [11] maps any variable-sized set of local features to a fixed-dimensional descriptor. Thus, the same network can aggregate a small set of training correspondences, a complete scene point cloud, or a segmented object instance without architectural modification. SALAD produces a descriptor $G \in \mathbb{R}^{K d_{\text{cl}}}$, where K is the number of learned clusters and d_{cl} is the dimensionality of each cluster vector.

Scene Retrieval For scene-level retrieval, the full scene point cloud is encoded using Utonia, projected into the shared embedding space, and aggregated with SALAD to obtain a point-cloud descriptor G^{point}. Let the scene be observed by V images, where the v-th image produces N_v DINOv2 patch tokens. We concatenate the patch tokens from all views before aggregation:

$$H^{\text{patch}} = \text{concat}\left(H_1^{\text{patch}}, \ldots, H_V^{\text{patch}}\right) \in \mathbb{R}^{(\sum_{v=1}^{V} N_v) \times d}.$$

These concatenated multi-view patch features are passed through the image projection layer and SALAD aggregator to produce the image-side scene descriptor G^{patch}. Retrieval is then performed by ranking database descriptors according to their cosine similarity with the query.

Instance Retrieval Instance-level retrieval restricts the inputs to a segmented object instance. The scene point cloud is replaced by the instance point cloud, and each query view is an RGB image together with the instance mask, of which only the patch tokens covered by the mask are used. For instance retrieval we use the local features directly instead of the SALAD descriptor: the projected patch features from all query views and the projected instance point features are each mean-pooled, and the two resulting descriptors are compared using cosine similarity. We use local features because they give better instance retrieval than the SALAD descriptor on both datasets (Table 4).

3.3 Implementation Details

CrossWeave is implemented in PyTorch [23] and trained on a single NVIDIA A6000 GPU using mixed-precision bfloat16 training. Training requires 17.1 GB of VRAM and about 21 GPU hours. We use the AdamW optimizer [16] with a learning rate of 10^{-3}, weight decay of 10^{-2}, and train for 30 epochs with a batch size of 16. During training, each sample contains 32 patch-point correspondences. DINOv2-L produces 1024-dimensional image patch features, and Utonia produces 1224-dimensional point features. These are projected into a shared 1024-dimensional embedding space using modality-specific projection matrices W_{patch} and W_{point}. The SALAD aggregator uses $K = 24$ cluster vectors with $d_{\text{cl}} = 128$, producing a final descriptor of dimension $K \times d_{\text{cl}} = 24 \times 128 = 3072$. We measure

inference time on a single NVIDIA A6000 GPU using 10 query images of resolution $3 \times 518 \times 518$ and a point cloud of $50{,}000$ points. The RGB forward pass (DINOv2 + SALAD) runs in 118.69 ± 1.87 ms and the point cloud forward pass (Utonia + SALAD) runs in 332.15 ± 4.52 ms, with a combined peak memory usage of 5.24 GB. In practice, the point cloud database descriptors are computed offline and cached, so only the RGB forward pass is required at query time.

4 Experiments

4.1 Datasets

ScanNet [5] ScanNet is a large-scale indoor dataset containing posed RGB-D frames from 707 scenes and 1513 scans, with multiple runs per scene. The dataset provides dense 3D meshes, with semantic and instance annotations for RGB frames and reconstructed meshes. We use the standard splits of 565 scenes / 1201 scans for training and 142 scenes / 312 scans for validation.

3RScan [32] 3RScan contains 1482 scans of 478 indoor scenes, with posed RGB-D frames, dense 3D meshes, and semantic and instance annotations. We use 384 scenes / 1177 scans for training and 46 scenes / 157 scans from the validation split for evaluation following [17].

Additional dataset details are provided in Appendix A.

4.2 Tasks and Metrics

Metrics For both scene-level and instance-level retrieval, we report *Top-k accuracy*, defined as the percentage of queries for which the correct scene or instance appears among the top-k retrieved results. For instance retrieval, we additionally report *Recall@k%* (R@k%), which measures the percentage of scenes for which at least k% of the query instances are correctly retrieved.

Scene Retrieval In the scene retrieval task, the query consists of 10 RGB views of a scene, selected following the protocol of CrossOver [27], and the database consists of 3D point clouds reconstructed from scanned scenes. A retrieval is considered correct if the model retrieves the point cloud corresponding to the queried scene. The database includes both original scans and re-scans, where the same environment may be captured again after objects have been displaced.

Instance Retrieval In the instance retrieval task, the query consists of multiple RGB views of an object instance, each with its instance mask. The model must retrieve the corresponding 3D object point cloud from the instance point clouds of the scene the query was taken from. The task therefore evaluates fine-grained object matching within a known environment. Top-k accuracy is computed within each scene and averaged over all validation scenes.

4.3 Baselines

Our primary baseline is CrossOver [27]. We evaluate two publicly released CrossOver checkpoints for each task: one trained on the same datasets as our method, and another checkpoint of the same model trained on additional data, which we denote CrossOver†. To ensure that performance differences are not caused by the choice of 3D encoder, we also retrain CrossOver on its original preprocessed training data after replacing its 3D backbone with Utonia [41]; we refer to this variant as *CrossOver (Updated)*. All methods are evaluated with CrossOver's evaluation protocol for a fair comparison.

5 Results

5.1 Cross-Modal Scene Retrieval

Scene retrieval results are reported in Table 1. CrossWeave consistently outperforms all baselines on both ScanNet and 3RScan across all retrieval ranks. On ScanNet, CrossWeave achieves a Top-1 recall of 61.54, improving over CrossOver (Updated) by 21.16 points. The gains remain strong at higher ranks, where CrossWeave reaches 99.36 Top-5 recall and perfect Top-10 recall. These results indicate that the correct scene is almost always ranked among the first few retrieved candidates.

Table 1: Comparison of scene retrieval performance on ScanNet and 3RScan. † Trained on additional datasets (ArkitScenes [2] + MultiScan [18]). **Best** and Second-best results.

	ScanNet			3RScan		
	Scene Recall			Scene Recall		
Method	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
Scene CrossOver	22.44	61.86	75.32	19.11	58.60	77.07
Scene CrossOver†	19.87	58.01	74.36	17.83	51.59	73.25
Scene CrossOver (Updated)	40.38	90.06	96.47	23.57	66.88	84.71
Ours	**61.54**	**99.36**	**100.00**	**46.50**	**85.99**	**92.36**

Table 2: Comparison of within-scene instance retrieval performance on ScanNet and 3RScan. † Trained on additional datasets (ArkitScenes [2] + MultiScan [18]). **Best** and Second-best results.

ScanNet

	Instance Retrieval		Scene-Level Recall		
Method	Top-1	Top-3	R@25%	R@50%	R@75%
Instance CrossOver	63.58	88.71	96.79	80.77	31.09
Instance CrossOver†	63.42	88.55	96.79	80.77	31.09
Instance CrossOver (Updated)	70.94	92.39	**99.36**	88.14	49.68
Ours	**78.77**	**94.05**	98.40	**88.78**	**67.95**

3RScan

	Instance Retrieval		Scene-Level Recall		
Method	Top-1	Top-3	R@25%	R@50%	R@75%
Instance CrossOver	58.88	84.24	90.45	71.97	28.03
Instance CrossOver†	57.09	84.58	89.81	63.06	24.84
Instance CrossOver (Updated)	66.80	90.85	97.45	87.90	35.03
Ours	**73.50**	**92.73**	**98.09**	**92.99**	**49.68**

The improvement is particularly meaningful because CrossOver (Updated) uses the same Utonia backbone. This suggests that the performance gain does not simply come from stronger 3D encoding, but from our training strategy based on sparse local patch-point supervision. By aligning local image patches with corresponding 3D points, CrossWeave learns descriptors that preserve fine-grained geometric and visual correspondences. This helps disambiguate scenes with similar global layouts, such as corridors, bedrooms, or repeated furniture arrangements, where purely global descriptors can collapse visually similar environments.

On 3RScan, CrossWeave shows the same trend. It achieves a Top-1 recall of 46.50, compared to 23.57 for CrossOver (Updated), corresponding to a 22.93 point improvement. CrossWeave also improves Top-5 recall from 66.88 to 85.99 and Top-10 recall from 84.71 to 92.36. These results demonstrate that local patch-point alignment generalizes across datasets and remains effective under the more challenging scene variations present in 3RScan. We also observe that CrossOver† does not improve over the base CrossOver model, suggesting that additional training data alone is insufficient without a training objective that explicitly encourages cross-modal local alignment.

Figure 2 shows a representative qualitative example on ScanNet. Given a set of query images, CrossWeave retrieves the correct 3D scene at rank 1, whereas both CrossOver baselines fail to retrieve the ground-truth scene within the top five results. This behavior is consistent across different scene categories: even when the query images cover only partial or localized views, CrossWeave produces descriptors that remain discriminative enough to separate visually similar scenes. Additional qualitative examples are provided in Appendix F, and failure cases are analyzed in Appendix G.

5.2 Cross-Modal Instance Retrieval

Instance retrieval results are reported in Table 2. Each query instance is retrieved from the instances of its own scene.

On ScanNet, CrossWeave achieves the best instance retrieval performance, with a Top-1 recall of 78.77 compared to 70.94 for CrossOver (Updated), an improvement of 7.83 points. CrossWeave also obtains the best Top-3 recall, improving from 92.39 to 94.05. Local patch-point supervision therefore improves fine-grained instance discrimination as well as scene matching.

Across scenes on ScanNet, the difference between methods becomes more pronounced at stricter recall thresholds. At R@25%, CrossWeave and CrossOver (Updated) are both near saturation, achieving 98.40 and 99.36, respectively. At R@50%, CrossWeave obtains the best result with 88.78, slightly outperforming CrossOver (Updated). The largest gain appears at R@75%, where CrossWeave reaches 67.95 compared to 49.68 for CrossOver (Updated), an improvement of 18.27 points. This

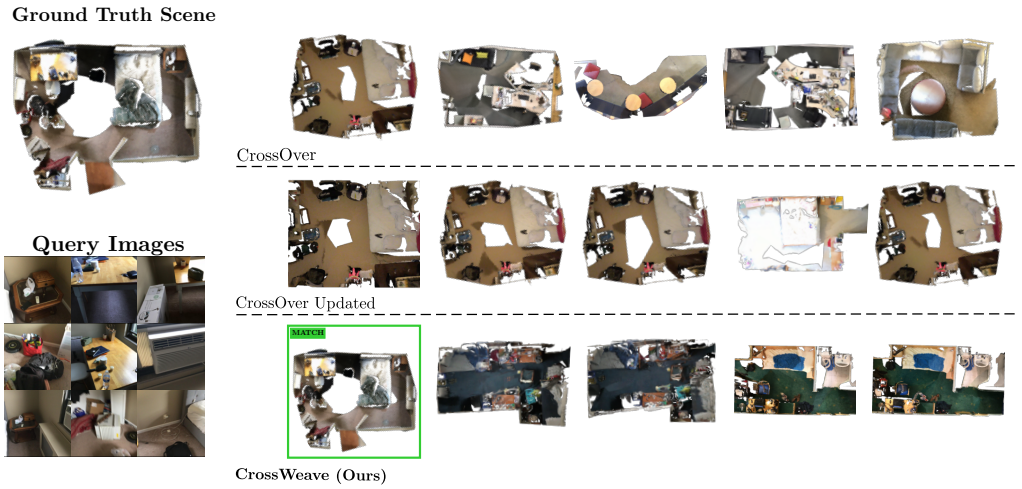


Figure 2: Qualitative comparison of cross-modal scene retrieval on ScanNet. Given a set of query RGB images, each model retrieves the most similar 3D scenes from the database. The ground-truth scene is shown on the left, and retrieved scenes are ordered by similarity from left to right. CrossOver and CrossOver (Updated) fail to retrieve the correct scene within the top five results, while CrossWeave retrieves it at rank 1. The remaining retrieved scenes are also semantically consistent, indicating that CrossWeave learns descriptors that capture both local cross-modal correspondences and global scene context.

suggests that CrossWeave is especially effective when the task requires retrieving a larger fraction of all correct instances, rather than only identifying the easiest matches.

The same pattern is observed on 3RScan. CrossWeave achieves the best instance retrieval performance, with a Top-1 recall of 73.50 and a Top-3 recall of 92.73, outperforming CrossOver (Updated) by 6.70 and 1.88 points, respectively. CrossWeave also obtains the best scene-level recall at all thresholds, reaching 98.09 at R@25%, 92.99 at R@50%, and 49.68 at R@75%. The improvement at R@75% is especially notable, where CrossWeave improves over CrossOver (Updated) by 14.65 points.

Overall, the instance retrieval results support the same conclusion as the scene retrieval experiments: sparse local patch-point alignment provides a stronger cross-modal training signal than instance-level supervision alone. While the baselines are often sufficient to retrieve a small number of easy matches, CrossWeave is more reliable under stricter retrieval criteria, where the model must recover a broader and more complete set of correct instances. Figure 3 shows two examples on ScanNet, where CrossWeave ranks the correct instance first among 16 and 26 candidates in the scene. Results with noisy 2D and 3D instance masks are reported in Appendix D.

5.3 Single View Scene Retrieval

Table 3 compares CrossWeave with SceneGraphLoc [19] and CrossOver [27] on 3RScan, where each query image is matched against N candidate scenes, following the evaluation protocol of SceneGraphLoc. SceneGraphLoc uses ground-truth scene graphs at inference, whereas CrossWeave uses only a single query image and the raw point cloud of each candidate scene. With 10 candidates, CrossWeave matches SceneGraphLoc at Top-1 (53.60) and improves Top-5 recall from 81.90 to 93.10 and Top-10 recall from 92.80 to 100.00. Recall drops for all methods when the pool grows to 50 candidates. In this setting, CrossWeave is 2.70 points behind SceneGraphLoc at Top-1 (27.50 vs. 30.20), but leads by 10.00 points at Top-5 and 15.30 points at Top-10. CrossWeave also outperforms CrossOver in every setting, by between 7.60 and 24.88 points. Since CrossWeave needs no scene graph annotations at inference, it can be deployed where such annotations are not available.

5.4 Ablation Studies

We ablate the main design choices of CrossWeave in Table 4. Specifically, we study the effect of the local patch-point alignment loss $\mathcal{L}_{p2p}$, the image-point-cloud aggregation loss $\mathcal{L}_{i2pc}$, the aggregation

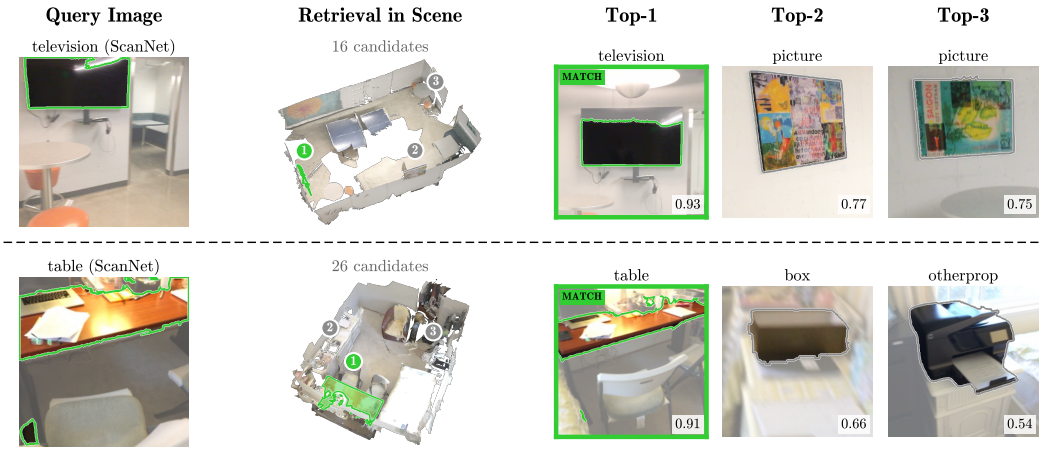


Figure 3: Qualitative instance retrieval on ScanNet. For each query view with its instance mask (left), CrossWeave ranks the instances of the same scene (middle) by cosine similarity. The Top-3 retrievals are shown with their similarity scores, and the correct match is outlined in green.

Table 3: Scene retrieval on 3RScan against SceneGraphLoc and CrossOver. Each query image is ranked against N candidate scenes; Top-k: percentage of queries with the correct scene in the top k. ‡ Uses ground-truth scene graphs at inference. **Best** and second-best results are highlighted.

	10 candidates			50 candidates		
Method	Top-1	Top-5	Top-10	Top-1	Top-5	Top-10
SceneGraphLoc‡ [19]	**53.60**	81.90	92.80	**30.20**	50.20	61.20
CrossOver [27]	46.00	77.97	90.58	18.69	39.16	51.62
Ours	**53.60**	**93.10**	**100.00**	27.50	**60.20**	**76.50**

used during training (mean pooling or SALAD), and the choice of retrieval features. We compare three feature variants: SALAD, which uses the aggregated descriptor produced by the SALAD module; LOCAL, which mean-pools the patch and point features; and HYBRID, which uses SALAD features for scene retrieval and LOCAL features for instance retrieval.

Local patch-point alignment improves both scene and instance retrieval Comparing the SALAD baseline trained only with $\mathcal{L}_{i2pc}$ against CrossWeave trained with both $\mathcal{L}_{i2pc}$ and $\mathcal{L}_{p2p}$ shows the benefit of local cross-modal supervision. On ScanNet, using SALAD features, CrossWeave improves scene Top-1 recall from 57.37 to 61.54, instance Top-1 recall from 73.33 to 77.04, and instance $R@75\%$ from 56.09 to 66.67. A similar trend is observed on 3RScan, where scene Top-1 recall improves from 43.95 to 46.50, instance Top-1 recall from 66.98 to 70.01, and instance $R@75\%$ from 38.22 to 43.31. These gains indicate that sparse local patch-point correspondences provide a stronger training signal than global image-point-cloud supervision alone, especially for fine-grained instance retrieval.

Local alignment alone is not sufficient for scene retrieval Training with only $\mathcal{L}_{p2p}$ and mean-pooled LOCAL features already gives strong instance retrieval, with 78.96 and 73.30 instance Top-1 recall on ScanNet and 3RScan. Scene retrieval, however, drops to 47.12 and 27.39 Top-1 recall, compared to 61.54 and 46.50 for the main model. Local alignment therefore learns discriminative object-level features, but does not by itself produce descriptors that separate scenes. Adding $\mathcal{L}_{i2pc}$ while keeping mean pooling improves scene Top-1 recall from 47.12 to 52.24 on ScanNet and from 27.39 to 30.57 on 3RScan, so the aggregated objective provides supervision that the local loss does not. Replacing mean pooling with SALAD further improves scene Top-1 recall to 61.54 and 46.50. Hence, both the aggregation objective and the aggregation module contribute to scene retrieval. Mean pooling with both losses gives the highest instance Top-1 recall (79.63 and 73.62), slightly above the main model (78.77 and 73.50), but its scene Top-1 recall is 9.30 and 15.93 points lower.

Table 4: Ablation of training losses, aggregation and retrieval features. *Agg.*: aggregation used during training. *Feats*: SALAD (aggregated descriptors), LOCAL (mean-pooled patch and point features), or HYBRID (SALAD for scene, LOCAL for instance retrieval). **Best** and second-best results are highlighted; the main model is shaded and marked with ⋆.

				ScanNet			3RScan		
				Scene	Instance		Scene	Instance	
Agg.	$\mathcal{L}_{i2pc}$	$\mathcal{L}_{p2p}$	**Feats**	Top-1	Top-1	R@75%	Top-1	Top-1	R@75%
Mean	✗	✓	LOCAL	47.12	78.96	70.51	27.39	73.30	47.13
Mean	✓	✓	LOCAL	52.24	**79.63**	**71.15**	30.57	**73.62**	49.04
SALAD	✓	✗	SALAD	57.37	73.33	56.09	43.95	66.98	38.22
SALAD	✓	✗	LOCAL	38.14	73.31	52.88	19.75	67.43	36.31
SALAD	✓	✓	SALAD	**61.54**	77.04	66.67	**46.50**	70.01	43.31
SALAD	✓	✓	LOCAL	48.40	78.77	67.95	28.66	73.50	**49.68**
SALAD	✓	✓	HYBRID⋆	**61.54**	78.77	67.95	**46.50**	73.50	**49.68**

Aggregated features are better suited for scene retrieval For scene-level retrieval, SALAD features consistently outperform LOCAL features. Without $\mathcal{L}_{p2p}$, the SALAD baseline achieves 57.37 Top-1 scene recall on ScanNet, compared to 38.14 when using local features. On 3RScan, the gap is even larger, with SALAD achieving 43.95 compared to 19.75 for LOCAL. The same pattern holds for CrossWeave: SALAD features achieve 61.54 scene Top-1 recall on ScanNet and 46.50 on 3RScan, whereas LOCAL features achieve only 48.40 and 28.66, respectively. This confirms that scene retrieval benefits from a parametric aggregation module over mean-pooled local features.

Local features are better suited for instance retrieval In contrast, instance retrieval benefits more from preserving local feature structure. For CrossWeave on ScanNet, using LOCAL features improves instance Top-1 recall from 77.04 to 78.77 and R@75% from 66.67 to 67.95 compared to SALAD features. On 3RScan, the improvement is larger: instance Top-1 recall increases from 70.01 to 73.50, and $R@75\%$ improves from 43.31 to 49.68. This suggests that aggregation can discard fine-grained spatial and geometric cues that are useful for distinguishing object instances, whereas local descriptors retain the detailed correspondence information required for accurate instance matching.

The hybrid feature strategy gives the best overall trade-off The HYBRID configuration combines the strengths of both feature types by using SALAD descriptors for scene retrieval and LOCAL for instance retrieval. This yields the best trade-off between scene and instance retrieval across both datasets, and we therefore use this configuration as our main model.

6 Conclusion

In this work, we introduced CrossWeave, a novel framework for cross-modal scene and instance retrieval that demonstrates the superiority of sparse, local feature alignment over traditional global scene encoding. By pairing a local patch-to-point contrastive objective with optimal transport-based aggregation, CrossWeave achieves state-of-the-art scene and instance retrieval on ScanNet and 3RScan with a single model, trained without scene- or instance-level contrastive objectives. CrossWeave provides a flexible alternative for bridging the modality gap in spatial reasoning tasks.

Acknowledgments and Disclosure of Funding

Aadith Warrier is supported by the Reliance Foundation Postgraduate Scholarship.

References

[1] Relja Arandjelovic, Petr Gronat, Akihiko Torii, Tomas Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5297–5307, 2016.

[2] Gilad Baruch, Zhuoyuan Chen, Afshin Dehghan, Tal Dimry, Yuri Feigin, Peter Fu, Thomas Gebauer, Brandon Joffe, Daniel Kurz, Arik Schwartz, et al. Arkitscenes: A diverse real-world dataset for 3d indoor scene understanding using mobile rgb-d data. *arXiv preprint arXiv:2111.08897*, 2021.

[3] Min Cao, Shiping Li, Juntao Li, Liqiang Nie, and Min Zhang. Image-text retrieval: A survey on recent research and development. *arXiv preprint arXiv:2203.14713*, 2022.

[4] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[5] Angela Dai, Angel X. Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proc. Computer Vision and Pattern Recognition (CVPR), IEEE*, 2017.

[6] Yu Duan, Yun Xiong, Yao Zhang, Yuwei Fu, and Yangyong Zhu. Hsgmp: Heterogeneous scene graph message passing for cross-modal retrieval. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*, pages 82–91, 2021.

[7] Y. Eldar, M. Lindenbaum, M. Porat, and Y.Y. Zeevi. The farthest point strategy for progressive image sampling. *IEEE Transactions on Image Processing*, 6(9):1305–1315, 1997. doi: 10.1109/83.623193.

[8] Yanglin Feng, Yang Qin, Dezhong Peng, Hongyuan Zhu, Xi Peng, and Peng Hu. Pointcloud-text matching: Benchmark dataset and baseline. *IEEE Transactions on Multimedia*, 2025.

[9] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15180–15190, June 2023.

[10] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, and Pheng-Ann Heng. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following, 2023. URL https://arxiv.org/abs/2309.00615.

[11] Sergio Izquierdo and Javier Civera. Optimal transport aggregation for visual place recognition. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 17658–17668, 2024.

[12] Justin Johnson, Ranjay Krishna, Michael Stark, Li-Jia Li, David Shamma, Michael Bernstein, and Li Fei-Fei. Image retrieval using scene graphs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3668–3678, 2015.

[13] Hervé Jégou, Matthijs Douze, Cordelia Schmid, and Patrick Pérez. Aggregating local descriptors into a compact image representation. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 3304–3311, 2010. doi: 10.1109/CVPR.2010.5540039.

[14] Nikhil Keetha, Avneesh Mishra, Jay Karhade, Krishna Murthy Jatavallabhula, Sebastian Scherer, Madhava Krishna, and Sourav Garg. Anyloc: Towards universal visual place recognition. *IEEE Robotics and Automation Letters*, 9(2):1286–1293, 2023.

[15] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 44860–44879. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/8c7304e77c832ddc70075dfee081ca6c-Paper-Conference.pdf.

[16] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[17] Haoyu Lu, Yuqi Huo, Mingyu Ding, Nanyi Fei, and Zhiwu Lu. Cross-modal contrastive learning for generalizable and efficient image-text retrieval. *Machine Intelligence Research*, 20(4):569–582, 2023.

[18] Yongsen Mao, Yiming Zhang, Hanxiao Jiang, Angel Chang, and Manolis Savva. Multiscan: Scalable rgbd scanning for 3d environments with articulated objects. *Advances in neural information processing systems*, 35:9058–9071, 2022.

[19] 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 *European Conference on Computer Vision*, pages 127–150. Springer, 2024.

[20] Kevin Musgrave, Serge J. Belongie, and Ser-Nam Lim. Pytorch metric learning. *ArXiv*, abs/2008.09164, 2020.

[21] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[22] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[23] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[24] Sai Shubodh Puligilla, Mohammad Omama, Husain Zaidi, Udit Singh Parihar, and Madhava Krishna. Lip-loc: Lidar image pretraining for cross-modal localization. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 939–948. IEEE, 2024.

[25] 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 Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[26] Junlong Ren and Hao Wang. Enhanced cross-modal 3d retrieval via tri-modal reconstruction. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2025.

[27] Sayan Deb Sarkar, Ondrej Miksik, Marc Pollefeys, Daniel Barath, and Iro Armeni. Crossover: 3d scene cross-modal alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8985–8994, June 2025.

[28] Konstantin Schall, Kai Uwe Barthel, Nico Hezel, and Klaus Jung. Optimizing clip models for image retrieval with maintained joint-embedding alignment. In *International Conference on Similarity Search and Applications*, pages 97–110. Springer, 2024.

[29] Chuan Tang, Xi Yang, Bojian Wu, Zhizhong Han, and Yi Chang. Parts2words: Learning joint embedding of point clouds and texts by bidirectional matching between parts and words. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6884–6893, 2023.

[30] Mikaela Angelina Uy and Gim Hee Lee. Pointnetvlad: Deep point cloud based retrieval for large-scale place recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[31] Kavisha Vidanapathirana, Milad Ramezani, Peyman Moghadam, Sridha Sridharan, and Clinton Fookes. Logg3d-net: Locally guided global descriptor learning for 3d place recognition. In *2022 International Conference on Robotics and Automation (ICRA)*, pages 2215–2221, 2022. doi: 10.1109/ICRA46639.2022.9811753.

[32] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Niessner. Rio: 3d object instance re-localization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

[33] Suping Wang, Fei Zhou, Ming Yang, Lei Shi, and Chaohong Tan. Sgg-mvar: Cross-modal retrieval with scene graph generation and multiview attribute relationship guidance. *IEEE Transactions on Computational Social Systems*, 12(5):3671–3683, 2025.

[34] Xiaoyang Wu, Li Jiang, Peng-Shuai Wang, Zhijian Liu, Xihui Liu, Yu Qiao, Wanli Ouyang, Tong He, and Hengshuang Zhao. Point transformer v3: Simpler faster stronger. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4840–4851, 2024.

[35] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1179–1189, June 2023.

[36] Le Xue, Ning Yu, Shu Zhang, Artemis Panagopoulou, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 27091–27101, June 2024.

[37] Chongjian Yuan, Jiarong Lin, Zheng Liu, Hairuo Wei, Xiaoping Hong, and Fu Zhang. Btc: A binary and triangle combined descriptor for 3-d place recognition. *IEEE Transactions on Robotics*, 40:1580–1599, 2024. doi: 10.1109/TRO.2024.3353076.

[38] Yihan Zeng, Chenhan Jiang, Jiageng Mao, Jianhua Han, Chaoqiang Ye, Qingqiu Huang, Dit-Yan Yeung, Zhen Yang, Xiaodan Liang, and Hang Xu. Clip2: Contrastive language-image-point pretraining from real-world point cloud data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15244–15253, June 2023.

[39] Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023.

[40] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Pointclip: Point cloud understanding by clip. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8552–8562, June 2022.

[41] Yujia Zhang, Xiaoyang Wu, Yunhan Yang, Xianzhe Fan, Han Li, Yuechen Zhang, Zehao Huang, Naiyan Wang, and Hengshuang Zhao. Utonia: Toward one encoder for all point clouds, 2026. URL https://arxiv.org/abs/2603.03283.

[42] Shuhang Zheng, Yixuan Li, Zhu Yu, Beinan Yu, Si-Yuan Cao, Minhang Wang, Jintao Xu, Rui Ai, Weihao Gu, Lun Luo, et al. I2p-rec: Recognizing images on large-scale point cloud maps through bird's eye view projections. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1395–1400. IEEE, 2023.

[43] Junsheng Zhou, Jinsheng Wang, Baorui Ma, Yu-Shen Liu, Tiejun Huang, and Xinlong Wang. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023.

[44] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyu Guo, Ziyao Zeng, Zipeng Qin, Shanghang Zhang, and Peng Gao. Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2639–2650, October 2023.

Appendix

A Data Pre-processing and Training Optimization Details

Both datasets are preprocessed by temporally subsampling the RGB-D frames: one in every 30 frames for ScanNet and one in every 5 frames for 3RScan, chosen according to the typical sequence length of each dataset. This yields 63,683 training images from ScanNet and 56,883 from 3RScan, for a total of 120,566 training images. For scene-level evaluation we use the 312 ScanNet and 157 3RScan validation scans. For instance-level evaluation we use 6,419 ScanNet and 3,274 3RScan validation instances.

To reduce dataloader overhead, projected pixel-point correspondences are precomputed and cached as `.pt` files, while FPS downsampling and patch-point pair generation are performed online. Utonia point features are also cached to avoid redundant forward passes over the same point clouds during training.

B Data Scaling

We study scene retrieval performance as a function of training data volume. The results reported in the main paper use 120,566 images from ScanNet and 3RScan as described in Appendix A. To study the effect of data scaling, we subsample the training set by uniform striding. As shown in Figure 4, CrossWeave is stable across data regimes: even at 10% of the full training set, the model outperforms the previous state of the art. We attribute the data efficiency of our model to two factors: feature-space FPS during patch-point pair selection, which ensures semantic diversity within each batch, and the use of strong pre-trained encoders for each modality, which provide rich local features from the outset. Together, these properties mean that even a small number of uniformly sampled images contains sufficient supervision signal to learn a discriminative cross-modal representation.

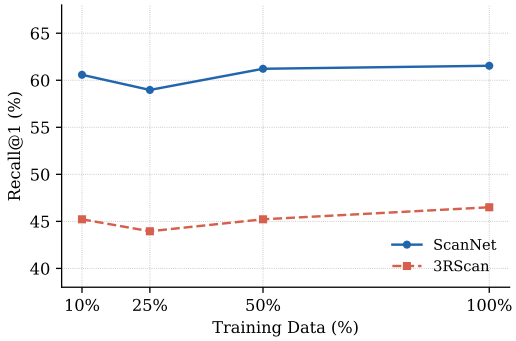


Figure 4: Scene retrieval performance (Top-1 recall) on ScanNet and 3RScan as a function of training data volume. Performance is stable across data regimes, with even 10% of the full training set sufficient to outperform the previous state of the art.

C Effect of Scaling Camera Views

Figure 5 reports scene retrieval performance as a function of the number of query camera views at inference. Top-1 recall improves from 45.83 to 64.10 on ScanNet and from 28.66 to 49.68 on 3RScan as the number of views increases from 3 to 30. Most of the gain comes between 3 and 10 views; beyond 10 views, recall still improves by up to 2.56 points on ScanNet and 4.46 points on 3RScan. This behavior is a direct consequence of the architecture: because the SALAD aggregator maps a variable number of input features into a fixed-dimensional descriptor, CrossWeave naturally incorporates additional views without any retraining or architectural modification. In contrast, CrossOver [27] averages per-frame features with equal weight, and its authors report that performance stabilizes after 10 frames, with additional frames giving only slight improvements.

CrossWeave continues to improve up to 25–30 views, since SALAD assigns patch features to clusters rather than averaging them. The ability to leverage more views at inference is particularly valuable in practical deployments, where a robot or AR device may capture varying numbers of observations depending on the environment.

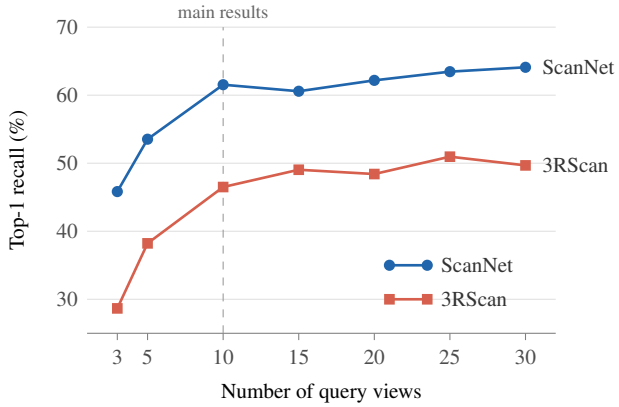


Figure 5: Scene retrieval performance (Top-1 recall) as a function of the number of query camera views. The main results use 10 views.

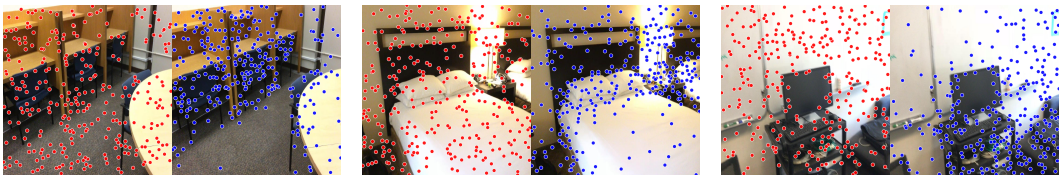
D Performance with Degraded Instance Masks

The instance retrieval results in Table 2 use ground-truth instance masks for both the query images and the database point clouds. In deployment, these masks come from a segmentation model and contain errors. We therefore evaluate the main model, with LOCAL features for instance retrieval, under noisy 2D masks, noisy 3D masks, and both. For 2D, we generate masks with MobileSAM [39], prompted with a single point sampled well inside the ground-truth mask to mimic a human click, and keep the predicted mask with the highest IoU to the ground truth. For 3D, we corrupt each ground-truth instance point cloud in two ways: points from neighboring background and objects bleed in at the instance boundary, and contiguous chunks of the instance are dropped. We set the corruption so that the IoU degradation is similar to that of the 2D masks. Results for each mask condition are reported in Table 5.

Table 5: Instance retrieval under noisy instance masks, using the main model with LOCAL features.

		Instance Retrieval		Scene-Level Recall		
Dataset	**Mask condition**	Top-1	Top-3	R@25%	R@50%	R@75%
ScanNet	GT	78.77	94.05	98.40	88.78	67.95
	2D noisy	75.75	93.02	97.44	88.14	62.50
	3D noisy	75.33	92.91	97.12	87.50	63.46
	2D + 3D noisy	72.29	91.86	97.12	85.90	55.45
3RScan	GT	73.50	92.73	98.09	92.99	49.68
	2D noisy	68.00	89.95	97.45	88.54	34.39
	3D noisy	65.65	88.79	96.82	84.08	32.48
	2D + 3D noisy	61.28	86.11	95.54	76.43	26.75

E FPS Sampling Visualizations



(a) ScanNet `scene0665_00` (b) ScanNet `scene0435_02` (c) ScanNet `scene0474_00`

Figure 6: Visualizations for the effect of FPS sampling. Images with red dots denote random sampling. Images with blue dots denote FPS sampled points. It can clearly be observed that FPS enables us to avoid over-sampling on the often visually dominant but geometrically uninformative areas in the images such as walls, floors and other flat surfaces, thereby boosting feature diversity within each training batch. Note: We have visualized a large number of pixel-point pairs in these images; in our training framework we only use 32 patch-point pairs. Refer to Section 3.1 for more details.

F Additional Visualizations for Qualitative Results

F.1 ScanNet Visualizations

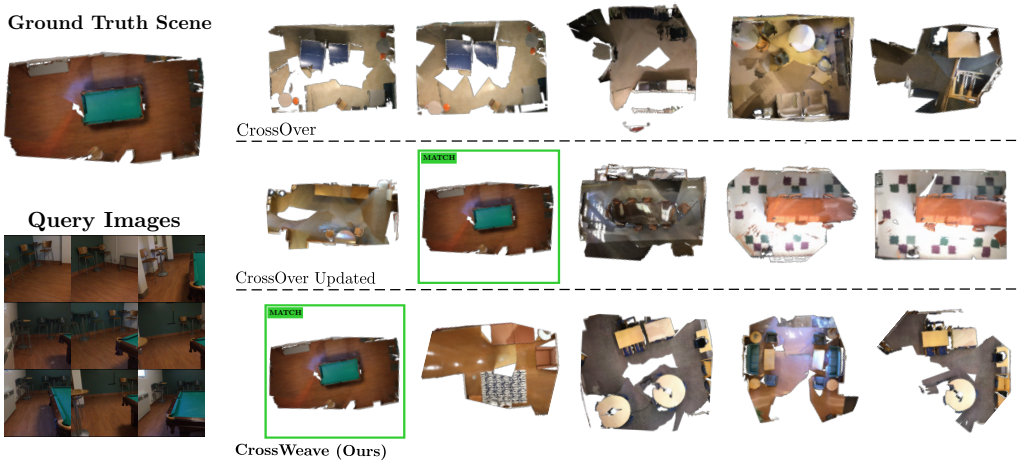


Figure 7: Comparisons of Top-5 retrievals on a ScanNet room scene. Retrievals are ordered from most to least similar moving left to right, with correct matches highlighted in green. CrossWeave successfully retrieves the correct scene at rank 1, while CrossOver (Updated) retrieves it at rank 2, and the base model of CrossOver fails within the top 5.



Figure 8: Additional qualitative results on a ScanNet office scene. CrossWeave retrieves the ground truth point cloud as its top match, whereas both CrossOver (Updated) and CrossOver only identify the correct scene at rank 5.

F.2 3RScan Visualizations

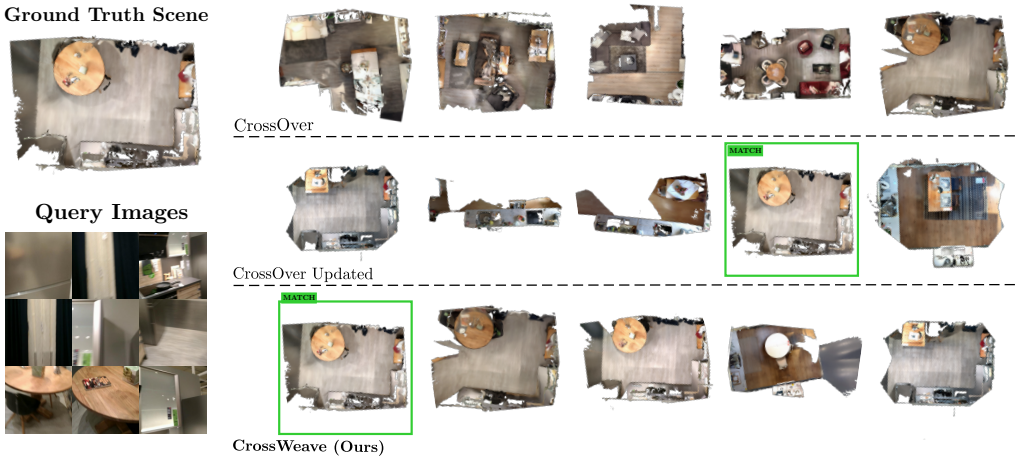


Figure 9: Qualitative results on a 3RScan dining area scene. CrossWeave successfully retrieves the ground truth point cloud as its top match, while CrossOver (Updated) ranks it fourth and the base CrossOver model fails to retrieve it within the top 5.

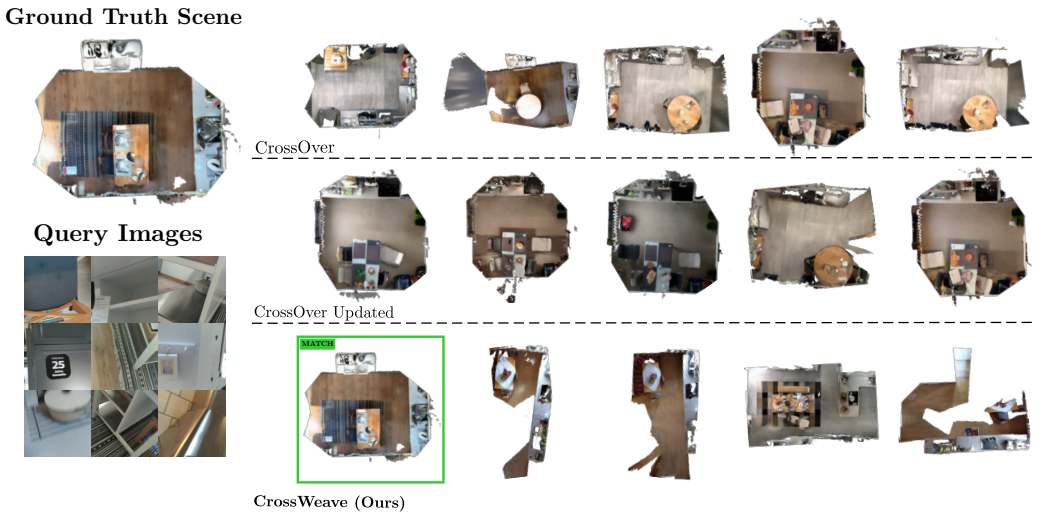


Figure 10: Retrieval results on a 3RScan kitchen and dining scene. CrossWeave demonstrates superior discriminative ability by retrieving the exact scene at rank 1 from highly localized query images, whereas both baseline models completely fail to place the correct match within their top 5 retrievals.

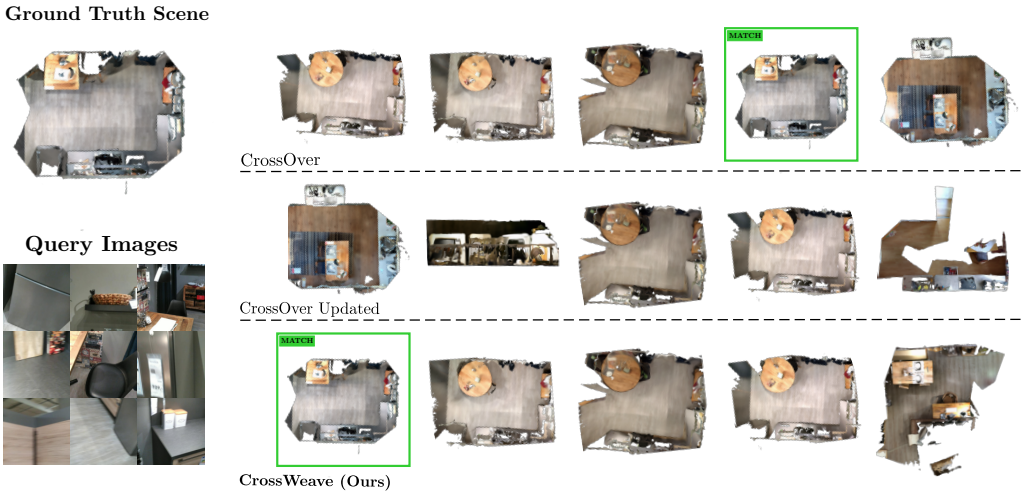


Figure 11: Additional results on a 3RScan kitchen scene. CrossWeave correctly identifies the target environment at rank 1, outperforming CrossOver, which places the match at rank 4, and CrossOver (Updated), which misses it entirely.

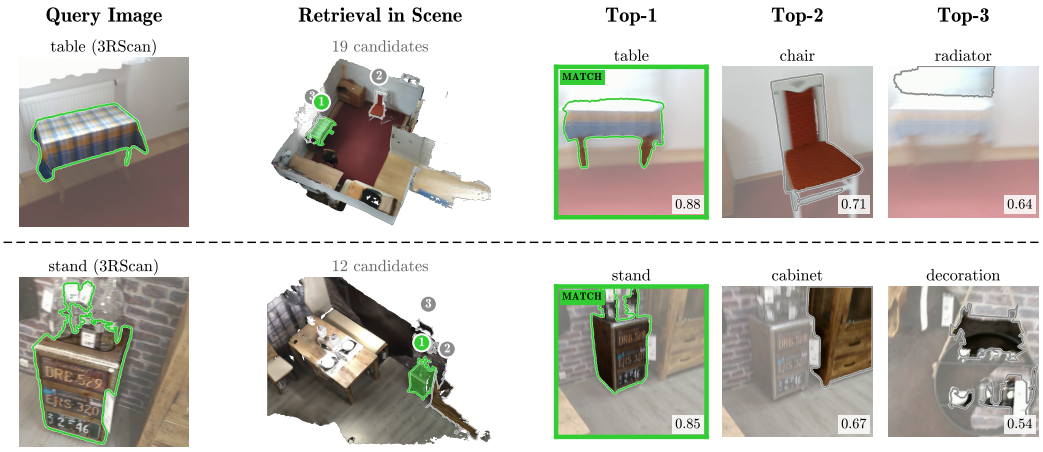


Figure 12: Qualitative instance retrieval on 3RScan. For each query view with its instance mask (left), CrossWeave ranks the instances of the same scene (middle) by cosine similarity. The Top-3 retrievals are shown with their similarity scores, and the correct match is outlined in green.

G Failure Case Analysis

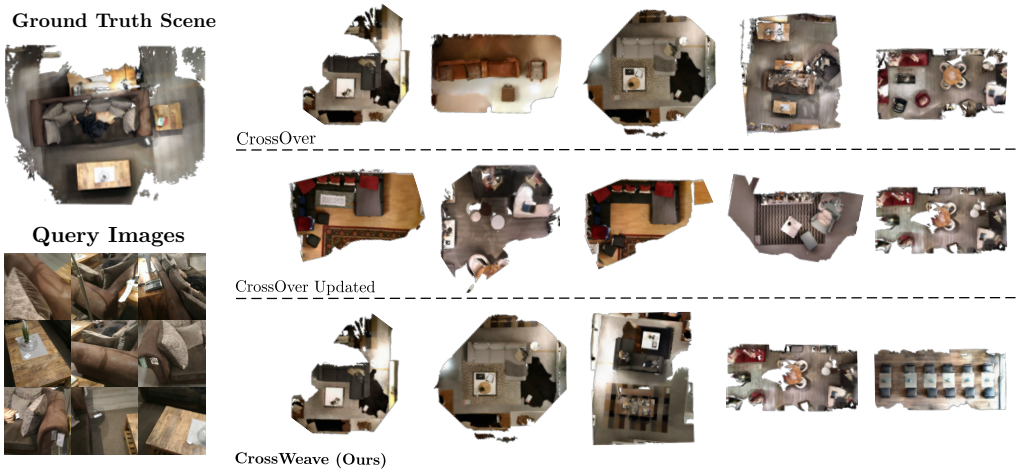


Figure 13: Failure case on a 3RScan living room scene. All three models miss the ground-truth scene in the top 5, as the retrieved scenes are visually and structurally similar to the query (a brown sectional sofa with a wooden coffee table). CrossWeave's top matches still share the furniture arrangement (sofas and central coffee tables), whereas CrossOver (Updated) retrieves structurally different scenes.

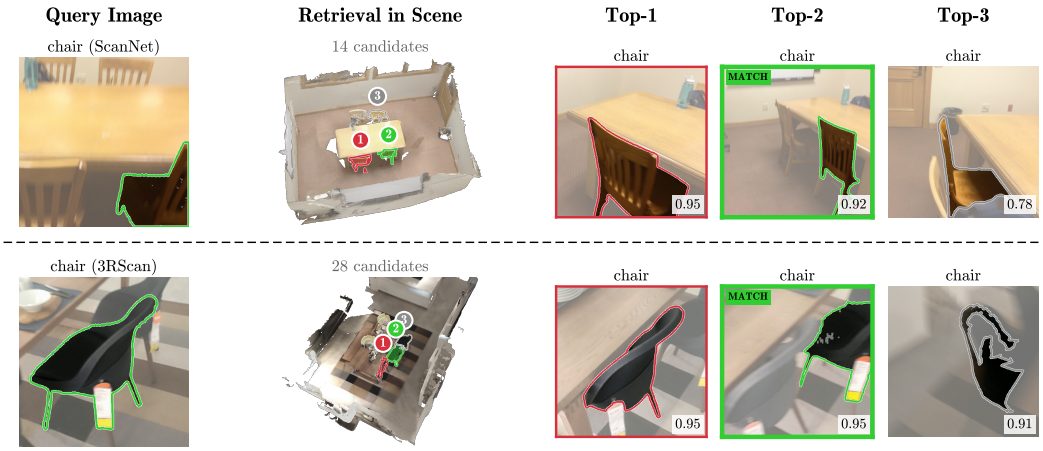


Figure 14: Instance retrieval failures on ScanNet and 3RScan. In both cases the query is a chair and the Top-1 retrieval (red) is another chair of the same type in the same scene, while the correct instance (green) is ranked second with a similar score.

H Limitations

Despite strong performance on most benchmarks, CrossWeave has several limitations. First, at inference, CrossWeave assumes that query images are captured within a single viewing session; queries that span multiple sessions are not explicitly handled. Second, the training correspondences are derived from the camera intrinsics, camera poses, depth rendered from the reconstructed mesh, and the reconstruction itself. Errors in calibration or pose, missing or inaccurate rendered depth, and incomplete or noisy reconstructions produce incorrect or missing patch-point pairs. We use the reconstructions and poses provided with ScanNet and 3RScan and do not study the effect of such errors. Finally, instance retrieval requires segmented instances, both as query masks and as database point clouds, making retrieval quality dependent on the upstream segmentation, which is outside the scope of this work. Appendix D shows that retrieval degrades under noisy 2D and 3D masks, most strongly at R@75%.

I Broader Impacts

CrossWeave enhances spatial intelligence for indoor robotics, autonomous navigation, and mixed-reality systems through efficient cross-modal localization. However, by relying on pre-trained foundation models (DINOv2 and Utonia), it risks inheriting and propagating their latent training biases. This could result in disparate retrieval performance across culturally or socioeconomically diverse architectural styles, causing the system to underperform in underrepresented environments. Real-world deployments must actively audit these underlying models to mitigate biased spatial reasoning.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and the introduction clearly state the core contributions of CrossWeave including the dual training losses and the SALAD based aggregation. The claims are backed by the results presented in Section 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

Justification: The limitations are discussed in Appendix H.

Guidelines:

- The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: This work is primarily empirical and does not introduce any formal theoretical results or proofs.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: All details necessary to fully reproduce the model including the exact hyperparameters are provided in Section 3.3 and appendix A and all the information regarding the datasets used is provided in Section 4.1.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The code for data pre-processing, training and evaluation, together with pre-trained checkpoints and instructions, is released at https://github.com/crossweave3d/crossweave. The datasets used (ScanNet and 3RScan) are publicly available.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Details about the dataset are provided in Section 4.1. Hyperparameters and the training configuration are detailed in Section 3.3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: We do not report error bars as the models are evaluated deterministically on standard fixed test splits of established Cross-modal retrieval benchmarks. Also we report results on recall metrics that are not amenable to error bar based analysis.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

Justification: These details are provided in Section 3.3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We use publicly available datasets widely used in existing research. There are no human-subject related concerns in this paper. We discuss some broader impact issues in Appendix I that might arise from bias present in the pre-trained models leveraged in this work.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Both impacts are discussed in the Appendix I.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The model that we release does not pose any such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The models and data are available under public licenses and all the datasets/models used have been appropriately cited.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The code and checkpoints are released at `https://github.com/crossweave3d/crossweave`.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.

- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No human subject or crowdsourcing is used for this work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: Since there is no research with human subjects, IRB approval is not applicable for this work.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used for development of any non-standard component of the core methodology of this work.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.