CrossOver: 3D Scene Cross-Modal Alignment

Sayan Deb Sarkar[1]   Ondrej Miksik[2]   Marc Pollefeys[2,3]   Daniel Barath[3,4]   Iro Armeni[1]

[1]Stanford University   [2]Microsoft Spatial AI Lab   [3]ETH Zurich   [4]HUN-REN SZTAKI

sayands.github.io/crossover

Abstract

Multi-modal 3D object understanding has gained significant attention, yet current approaches often assume complete data availability and rigid alignment across all modalities. We present CrossOver, a novel framework for cross-modal 3D scene understanding via flexible, scene-level modality alignment. Unlike traditional methods that require aligned modality data for every object instance, CrossOver learns a unified, modality-agnostic embedding space for scenes by aligning modalities – RGB images, point clouds, CAD models, floorplans, and text descriptions – with relaxed constraints and without explicit object semantics. Leveraging dimensionality-specific encoders, a multi-stage training pipeline, and emergent cross-modal behaviors, CrossOver supports robust scene retrieval and object localization, even with missing modalities. Evaluations on Scan-Net and 3RScan datasets show its superior performance across diverse metrics, highlighting CrossOver's adaptability for real-world applications in 3D scene understanding.

1. Introduction

In recent years, the need to align and transfer information across modalities has grown substantially, especially for tasks involving complex 3D environments. Such a capability enables knowledge and experience transfer across modalities. For example, knowing the layout of kitchens in computer-aided design (CAD) format will provide guidance on how to build a new kitchen, such that it follows the layout of the most similar CAD floorplan.

Current multi-modal approaches tackle 3D data alignment of individual objects across modalities [18, 42, 43, 45], *without including and considering scene context*, making them challenging to extend effectively for scene-level understanding. These methods typically assume fully aligned, consistent datasets, where each modality is perfectly corresponding to all others for each object. However, real-world scenarios rarely provide such complete modality pairings. For example, a video of a room and its CAD model might share some spatial alignment but differ in
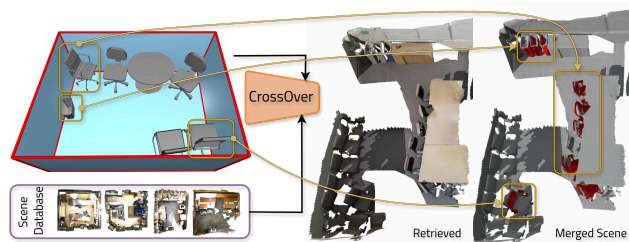

Figure 1. *CrossOver* **is a cross-modal alignment method for 3D scenes that learns a unified, modality-agnostic embedding space, enabling a range of tasks.** For example, given the 3D CAD model of a query scene and a database of reconstructed point clouds, CrossOver can retrieve the closest matching point cloud and, if object instances are known, it can identify the individual locations of furniture CAD models with matched instances in the retrieved point cloud, using brute-force alignment. This capability has direct applications in virtual and augmented reality.

data characteristics and object instances (hereby referred to as *instances*) represented in the data (e.g., some instances could be missing in one modality, which is common between real-world scenes and their CAD models). Also, achieving consistent instance segmentation across modalities is nearly impossible in practice. Thus, these approaches struggle when certain modalities are missing or incomplete, limiting their flexibility in practical applications [5].

We address the inherent limitations of strict object-level modality alignment by introducing a *flexible scene-level* modality alignment approach that operates without prior information during inference (*e.g.*, semantic instance segmentation), unlike the current methods [34, 41]. Our method, namely *CrossOver* (Fig. 1), enables the learning of cross-modal behaviors and relationships, such as identifying similar objects or scenes across different modalities, like the virtual CAD scene based on a video of a real room. This capability extends beyond instance-level matching towards a *unified, modality-agnostic understanding* that supports seamless cross-modal interactions at the scene level.

CrossOver focuses on aligning five key scene modalities—RGB images, real-world point clouds,

CAD models, floorplan images, and text descriptions, in the *feature* space—going beyond the RGB-PC-Text triplets of prior work. Importantly, it is designed with the assumption that not all modalities are available for every data point. By employing a flexible training strategy, we allow CrossOver to leverage any available modality during training, without requiring fully aligned data across all modalities. This approach enables our encoders to learn emergent modality alignments, supporting cross-modal traversals even in cases with missing data. Our work is grounded in three key contributions:

- **Dimensionality-Specific Encoders:** We introduce 1D, 2D, and 3D encoders tailored to each modality's dimensionality, removing the need for explicit 3D scene graphs or semantic labels during inference. This optimizes feature extraction for each modality and avoids reliance on consistent semantics, which is often hard to obtain.
- **Three-Stage Training Pipeline:** Our pipeline progressively builds a modality-agnostic embedding space. First, object-level embeddings capture fine-grained modality relationships. Next, scene-level training develops unified scene representations without requiring all object pairs to align. Finally, dimensionality-specific encoders create semantic-free cross-modal embeddings.
- **Emergent Cross-Modal Behavior:** CrossOver learns emergent modality behavior, despite not being explicitly trained on all pairwise modalities. It recognizes, *e.g.*, that $Scene_i$ in the image modality corresponds to $Scene_i$ in the floorplan modality or its point cloud to the text one, without these modality pairs being present in training.

This unified, modality-agnostic embedding space enables diverse tasks such as object localization and cross-modal scene retrieval, offering a flexible, scalable solution for real-world data that may lack complete pairings.

2. Related Work

Multi-modal Representation Learning aims to bridge data modalities by learning shared embeddings for cross-modal understanding and retrieval. A seminal work in this area is CLIP [32], which popularized the contrastive training objective to learn a joint image-text embedding space. This framework has been extended to various tasks, such as video retrieval [27], unified vision-language modeling [26], and cross-modal alignment [16, 28]. In the 3D domain, PointCLIP [45] applied CLIP to point clouds by projecting them into multi-view depth maps, leveraging pretrained 2D knowledge. Subsequent research has focused on multi-modality alignment, *e.g.* ImageBind [17] aligns six modalities in the 2D domain and shows the power of such representation for generative tasks. In 3D, ULIP [42] and its successor ULIP-2 [43] aim to learn unified representations among images, texts, and point clouds. Point-Bind [18] extends ImageBind [17] to 3D by aligning specific pairs of modal-

ities using an InfoNCE loss [30]. While these methods effectively capture object-level data, they struggle to differentiate similar instances within a scene, primarily focusing on isolated objects rather than complex scenes. Experiments in Section 4 demonstrate this limitation.

A common limitation of these approaches is the assumption of *perfect modality alignments* and *complete data* for each instance, often relying on datasets like ShapeNet55 [6]. This assumption is impractical for real-world scenarios where data is often incomplete or not well-matched due to occlusions, dynamic changes, sensor limitations, or capture errors, such as in construction sites or robot navigation. Our work, CrossOver, addresses these challenges using real-world datasets consisting of incomplete point clouds and noisy images captured using affordable sensors. Unlike prior methods, we do not require perfect modality alignments or complete data (e.g., point clouds).

3D Scene Understanding has driven extensive work on text-to-image and point cloud based instance localization and alignment within large maps [2, 14, 22]. Techniques like NetVLAD [2] and CamNet [14] enable place recognition and image-based localization by extracting global image descriptors. Recent work has leveraged 3D scene graphs for enhanced scene understanding [3, 21, 33], with methods like SGAligner [34] and SG-PGM [41] facilitating scene alignment through 3D scene graph matching. For dynamic instance matching across long-term sparse environments, LivingScenes [47] parses an evolving 3D environment with an object-centric formulation. For cross-modal retrieval, approaches like ScanRefer [7] and ReferIt3D [1] localize objects in 3D scenes via natural language but rely on detailed annotations and fixed modality pairs. Methods like 3DSSG [39] and "Where Am I" [8] extend scene retrieval across images and natural language using 3D scene graphs, yet they depend heavily on semantic annotations. SceneGraphLoc [29] performs image-to-scene-graph matching, using semantic information. Our approach diverges from these by removing the need for semantics or explicit scene graphs, instead leveraging dimensionality-specific encoders and modality-agnostic embeddings for scene understanding without prior semantic knowledge.

Handling Missing Modalities and noisy data is a key challenge in multi-modal learning [5]. Traditional approaches often assume full data availability, limiting their real-world applicability. Some methods address missing data through modality imputation or robust models [37, 40]. Baltrusaitis *et al*. [5] highlight that many methods lack flexibility for incomplete or noisy data. Our framework tackles this by allowing independent mapping of each modality into a shared embedding space, enabling flexible cross-modal interactions in unstructured environments with sparse or unaligned data. Furthermore, emergent behavior in multi-modal models, such as generalizing and inferring relationships beyond
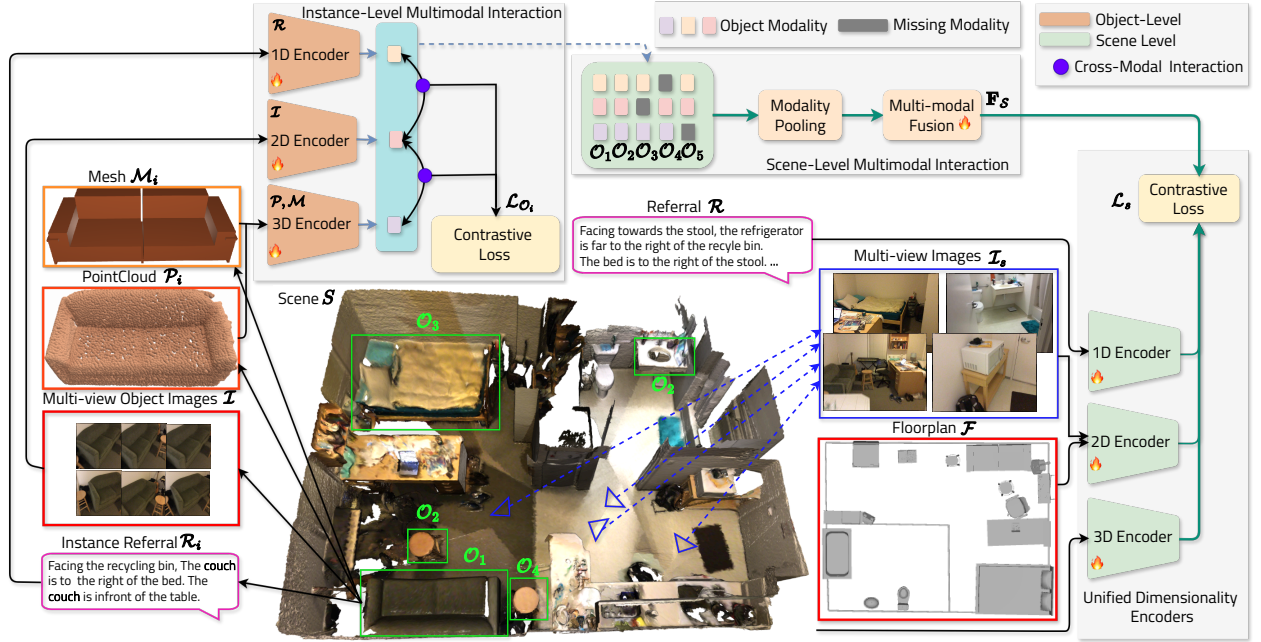
Figure 2. **Overview of CrossOver.** Given a scene $\mathcal{S}$ and its instances $\mathcal{O}_i$ represented across different modalities $\mathcal{I}, \mathcal{P}, \mathcal{M}, \mathcal{R}, \mathcal{F}$, the goal is to align all modalities within a shared embedding space. The *Instance-Level Multimodal Interaction* module captures modality interactions at the instance level within the context of a scene. This is further enhanced by the *Scene-Level Multimodal Interaction* module, which jointly processes all instances to represent the scene with a single feature vector $\mathcal{F}_\mathcal{S}$. The *Unified Dimensionality Encoders* eliminate dependency on precise semantic instance information by learning to process each scene modality independently while interacting with $\mathcal{F}_\mathcal{S}$.

training data [17, 32], are promoted by structuring training around image embeddings as a common representation. By mapping other modalities into this shared space, CrossOver fosters organic cross-modal relationships, enabling unified understanding across diverse data types.

3. Method

Given a 3D scene $\mathcal{S}$ represented by various modalities, denoted as $\mathcal{Q} = \{\mathcal{I}, \mathcal{P}, \mathcal{M}, \mathcal{R}, \mathcal{F}\}$, our objective is to develop a unified, modality-agnostic representation that maps independent modalities capturing the same 3D scene to a common point in the embedding space. Here, $\mathcal{I}$ is a set of RGB images, $\mathcal{P}$ is a real-world reconstruction as a point cloud, $\mathcal{M}$ is a digital mesh representation from computer aided design (CAD), $\mathcal{R}$ is textual data describing $\mathcal{S}$ within its surroundings, and $\mathcal{F}$ is a rasterized floorplan.

Our proposed framework facilitates robust interactions across different modalities at both the comprising instances and scene levels, enhancing the multi-modal (*e.g.*, pointcloud $\mathcal{P}$ and floorplan $\mathcal{F}$) and same modal (e.g., textual data $\mathcal{R}$) understanding of 3D environments. We structure the development of the embedding space progressively, beginning with instance-level multi-modal interactions and culminating in scene-level multi-modal interactions without requiring prior knowledge, such as semantic information about constituent instances. An overview of CrossOver is shown

in Fig. 2. To demonstrate the capabilities of this unified, modality-agnostic embedding space, we evaluate:

1. **Cross-modal instance retrieval**: Given an observed modality $\mathcal{Q}_j$ of a query instance $\mathcal{O}_i$ in a scene $\mathcal{S}$ (*e.g.*, mesh $\mathcal{M}$ or pointcloud $\mathcal{P}$), we aim to retrieve any other modality $\mathcal{Q}_k$ representing $\mathcal{O}_i$ within $\mathcal{S}$.
2. **Cross-modal scene retrieval**: Given a scene $\mathcal{S}_i$ represented by modality $\mathcal{Q}_j$ (*e.g.*, image $\mathcal{I}$ or floorplan $\mathcal{F}$), we aim to retrieve another modality $\mathcal{Q}_k$ representing $\mathcal{S}_i$.

3.1. Instance-Level Multi-Modal Interactions

First, we describe the pipeline used for learning a multimodal embedding space for independent instances. This will provide a basis for the scene-level embeddings. We process each of the 1D ($\mathcal{R}$), 2D ($\mathcal{I}$), and 3D ($\mathcal{P}$ and $\mathcal{M}$) instance modalities with corresponding encoders[1]:

1D Encoder. An instance $\mathcal{O}_i$ can be represented by its textual context in a scene $\mathcal{S}$, using descriptions like *"The chair is in front of the lamp"* and *"The chair is left of the table"*. We term these descriptions as *object referrals* [20] and encode each referral as $f_{ij}^\mathcal{R}$ using the pre-trained text encoder BLIP [23], where i is the instance of interest (*e.g.*, chair) and j represents another instance in the scene (*e.g.*, lamp, table, or another chair). Practically, we collect k object re-

---

[1]The $\mathcal{F}$ modality is not used when learning an instance-level embedding since there is no notion of a floorplan in this scenario.

ferrals per instance, resulting in $F_i^{\mathcal{R}} = \{f_{i1}^{\mathcal{R}}, \ldots, f_{ik}^{\mathcal{R}}\}$. To create a single feature vector $f_i^{\mathcal{R}}$ representing the instance's context, we apply average pooling over $F_i^{\mathcal{R}}$.

2D Encoder. Given a collection $I_{\mathcal{S}}$ of images capturing a scene $\mathcal{S}$, we integrate multi-view and per-view multi-level visual embeddings for each $\mathcal{O}_i$ to encode $f_i^{\mathcal{I}}$. Inspired by [29], for each $\mathcal{O}_i$, we select the top K_{view} defined by largest visibility of $\mathcal{O}_i$ among $I_{\mathcal{S}}$ and calculate multi-level bounding boxes around $\mathcal{O}_i$ $\{b_{v,l} \mid l \in [0, L)\}$ within each view v. A pre-trained DinoV2 [12, 31] encoder processes the image crops defined by $b_{v,l}$ to give us the [CLS] tokens per crop [44]. Subsequent average pooling operations aggregate these tokens into a singular feature vector $f_i^{\mathcal{I}}$. In contrast to [29], we do not assume available camera poses.

3D Encoder. Given instance $\mathcal{O}_i$ and its corresponding real-world *point cloud* $\mathcal{P}_i$ and *shape mesh* $\mathcal{M}_i$, we extract instance features $\bar{f}_i^{\mathcal{P}}$ and $f_i^{\mathcal{M}}$ using a pretrained I2PMAE [46] point cloud encoder. Importantly, we do not utilize the semantic class [20, 48] of $\mathcal{O}_i$ in these operations. We concatenate the 3D location of $\mathcal{P}_i$ and $\mathcal{M}_i$ to $\bar{f}_i^{\mathcal{P}}$ and $f_i^{\mathcal{M}}$, respectively, to form the instance tokens $\hat{f}_i^{\mathcal{P}}$ and $\hat{f}_i^{\mathcal{M}}$. To introduce partial scene-level reasoning, we incorporate interactions between instances by integrating the instance tokens and encoding the pairwise spatial relationships of an instance with all others in $\mathcal{S}$ within a transformer network. Similar to [20], we employ spatial-attention-based transformers, following [9, 48], to generate $f_i^{\mathcal{P}}$ and $f_i^{\mathcal{M}}$. Details about the 3D location and spatial relationships are in Supp. For the mesh modality $\mathcal{M}$, we sample points on the mesh surface to enable input to a point cloud encoder. We encode neither the 3D location nor the spatial pairwise relation among instances, as we do not assume that the meshes are aligned with the scene geometry.

All pre-trained encoders, which are frozen during training, are followed by trainable projection layers. During training, after encoding each modality, we apply a contrastive loss to enforce alignment of modality features within a joint embedding space. Unlike prior work that requires full data modality alignment [18, 43] or semantic scene graph [29, 34], CrossOver accommodates the practical challenge that not all modalities may always be available by not requiring the presence of all modalities simultaneously. Instead, it aligns all other modality embeddings with image space $\mathcal{I}$. The loss function can be defined as:

$$\mathcal{L}_{\mathcal{O}_i} = \mathcal{L}_{f^I, f^{\mathcal{P}}} + \mathcal{L}_{f^I, f^{\mathcal{M}}} + \mathcal{L}_{f^I, f^{\mathcal{R}}}. \tag{1}$$

During *training*, CrossOver requires a base modality for every instance, to align other modalities with its feature space. We choose images $\mathcal{I}$ as the base modality due to their availability and strong encoder priors, though any supported modality can serve this role. Crucially, *no modality availability assumptions are made during inference*, allowing any query-target modality pair. Our experiments (see

Supp.) show that aligning to a single reference modality, rather than using all pairwise combinations as in prior work, improves performance.

3.2. Scene-Level Multi-Modal Interactions

We distill knowledge from instance-level modality encoders to scene-level encoders, allowing us to leverage instance-based insights during training and enabling scene-level retrieval at inference without relying on 3D scene graphs or semantic instance information across modalities.

Multi-modal Scene Fusion. Given the instance features $f_i^{\mathcal{R}}$, $f_i^{\mathcal{I}}$, $f_i^{\mathcal{P}}$, and $f_i^{\mathcal{M}}$ for each instance $\mathcal{O}_i$ in scene $\mathcal{S}$, we compute each of the scene level features $f^{\mathcal{R}}$, $f^{\mathcal{I}}$, $f^{\mathcal{P}}$, and $f^{\mathcal{M}}$ by first applying average pooling per modality to the features of all instances in $\mathcal{S}$. We then perform a weighted fusion of these pooled features to learn a fixed-size multi-modal embedding $\mathbf{F}_{\mathcal{S}}$:

$$\mathbf{F}_{\mathcal{S}} = \sum_{q \in \mathcal{Q}} \left[\frac{\exp(w_q)}{\sum_{j \in \mathcal{Q} \setminus q} \exp(w_j)} f^q \right], \tag{2}$$

where $j, q \in \mathcal{Q}$, w_q and w_j are modality-wise trainable attention weights. We use an MLP head to project the dimensionality to our final representation space, resulting in an embedding that serves as a unified scene representation, capturing interactions across all modalities. In practice, this representation is flexible, adapting to data availability and specifically to any missing modalities.

3.3. Unified Dimensionality Encoders

The above scene-level encoder provides a unified, modality-agnostic embedding space; however, it requires semantic instance information consistent across modalities during inference, which is challenging to obtain in practice. To eliminate this need, we design a single encoder per modality dimensionality (*i.e.* 1D, 2D, and 3D) that directly processes raw data without needing additional information. Moreover, our experiments (Supp.) show that the scene-level encoder needs all modalities at inference to perform reasonably.

1D Encoder. Similar to Sec. 3.1, we use *object referrals* to describe scene context [48]. We randomly sample $t = 10$ referrals per scene and use a text encoder to form $\mathbf{F}_{1D}$.

2D Encoder. Here, we consider both RGB and floorplan images. The floorplan $\mathcal{F}$ is represented as a top-view orthographic projection image of the $3D$ layout with geometrically aligned shape meshes for furniture instances. Since a scene can be captured with multiple RGB images $\mathcal{I}_{\mathcal{S}}$, we employ a naive key-frame selection strategy to sample $N = 10$ multi-view images (see Supp.). We process the images using a DinoV2 [31] encoder and concatenate the output [CLS] token and aggregated patch embeddings to form $\mathbf{F}_{2D}^i, i \in N$. We pass each $\mathbf{F}_{2D}^i$ via an MLP projection head and apply average pooling to generate F_{2D}. In practice, we use the same encoder with shared weights for

both RGB images $\mathcal{I}_S$ and floorplan $\mathcal{F}$; *i.e.*, inputs are not distinguished between RGB and floorplan during training. This is the first use of the floorplan modality in CrossOver and there is *no pairwise modality interaction* during training between it and the image modality, unlike other modalities.

3D Encoder. We utilize a sparse convolutional architecture with a residual network as the encoder, built with the Minkowski Engine [10]. Given an input point cloud $P \in \mathbb{R}^{N \times 3}$ containing N points, it is first quantized into M_0 voxels represented as $V \in \mathbb{R}^{M_0 \times 3}$. The model then produces a full-resolution output feature map $\mathbf{F}_{3D} \in \mathbb{R}^{M_0 \times D}$.

The goal is to align each of the unified dimensionality encoders with the scene-level multi-modal encoder. The loss function for unified training becomes:

$$\mathcal{L}_s = \alpha\mathcal{L}_{\mathbf{F}_S, \mathbf{F}_{1D}} + \beta\mathcal{L}_{\mathbf{F}_S, \mathbf{F}_{2D}} + \gamma\mathcal{L}_{\mathbf{F}_S, \mathbf{F}_{3D}}, \quad (3)$$

where, α, β, and γ are learnable hyper-parameters.

Thus, our combined loss is as follows:

$$\mathcal{L} = \mathcal{L}_s + \sum_{\mathcal{O}_i \in \mathcal{S}} \mathcal{L}_{\mathcal{O}_i} \quad (4)$$

3.4. Loss Definition

Given $q = G(\mathcal{Q}^m{}_i)$ and $k = H(\mathcal{Q}^n{}_i)$, $i \in \mathcal{B}$, two different encoder outputs for modalities $\mathcal{Q}^m$ and $\mathcal{Q}^n$ in minibatch $\mathcal{B}$, we use a contrastive loss similar to [17]:

$$\mathcal{L}_{q,k} = -\log \frac{exp(q_i^T k_i / \tau)}{exp(q_i^T k_i / \tau) + \sum_{j \neq i} exp(q_i^T k_j / \tau)}. \quad (5)$$

Here, τ is a learnable temperature parameter, to modulate similarity between positive pairs. We consider every example $j \neq i$ in a minibatch $\mathcal{B}$ as a negative example. In practice, we use a symmetric loss for better convergence: $\mathcal{L}_{q,k} + \mathcal{L}_{k,q}$. Although we pair each modality with the most prevalent one (*i.e.*, $\mathcal{I}$) to avoid the need for fully aligned modalities per data point during training, there are cases where not all modality pairs are available for a given data point. To enhance CrossOver's flexibility, we account for these scenarios by masking the corresponding loss term for any unavailable modality pairs.

3.5. Inference

After training CrossOver with the loss objective defined in Eq. 4, we use the embedding feature vectors for retrieval tasks. Given a scene S containing $\mathcal{O} = \{\mathcal{O}_i\}$ instances each represented by one or more modalities from $\mathcal{Q}$, we use our instance-level multi-modal encoders to perform cross-modal retrieval. Given $\mathcal{O}_i$ in query modality $\mathcal{Q}_j$ and all other instances in target modality $\mathcal{Q}_k$, the goal is to retrieve the $\mathcal{O}_i$ in $\mathcal{Q}_k$. For scene retrieval, we apply a similar approach using our unified dimensionality encoders, except that instead of instances, we retrieve entire scenes. A schematic diagram for one modality pair is shown in Fig. 3.
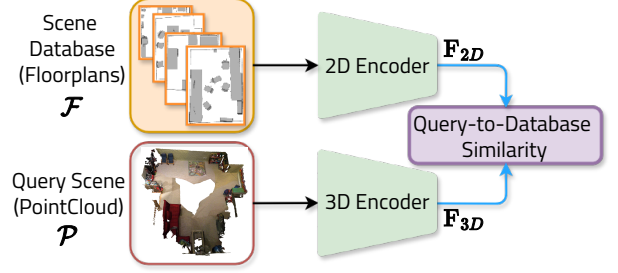

Figure 3. **Cross-modal Scene Retrieval Inference Pipeline.** Given a query modality ($\mathcal{P}$) that represents a scene, we obtain with the corresponding dimensionality encoder its feature vector ($\mathcal{F}_{3D}$) in the shared cross-modal embedding space. We identify the closest feature vector ($\mathcal{F}_{2D}$) in the target modality ($\mathcal{F}$) and retrieve the corresponding scene from a database of scenes in $\mathcal{F}$.

4. Experiments

Datasets. We train and evaluate CrossOver on ScanNet [11] and 3RScan [38]. We choose ScanNet for providing comprehensive coverage of all modalities, and 3RScan for including more data on temporal scenes. For both, we use the *object referrals* from SceneVerse [20], which is a million-scale 3D vision-language dataset with 68K 3D indoor scenes comprising indoor scene understanding datasets and $2.5M$ vision-language pairs. In all evaluations, we use a model trained across all datasets (details in Supp.).

ScanNet [11] is an RGB-D video dataset containing 2.5 million views in more than 1500 scenes, annotated with 3D camera poses, surface reconstructions, and instance-level semantic segmentation; we obtain images and 3D point clouds. For mesh $\mathcal{M}$ and floorplan $\mathcal{F}$, we use the Scan2CAD [4] dataset, which provides annotated keypoint pairs between CAD models from ShapeNet [6] and their counterpart objects in the scans. **3RScan [38]** benchmarks instance relocalization, featuring 1428 RGB-D sequences across 478 indoor scenes, including rescans of the latter after object relocation. It provides annotated 2D and 3D instance segmentation, camera trajectories, and reconstructed scan meshes. We obtain images and point clouds.

Evaluation Metrics. We assess the quality of our representation by quantifying its ability to identify the same instance $\mathcal{O}_i$ or scene $\mathcal{S}_i$ across modalities, $\mathcal{Q}_j$ and $\mathcal{Q}_k$. Extending image feature matching evaluation [25, 35], we compute the *instance matching recall* as the ratio of correctly identified $\mathcal{O}_i$ matches, given a database of instances. Additionally, we evaluate *scene-level (instance) matching recall* at thresholds of 25%, 50%, and 75%, indicating how many objects from a scene in modality $\mathcal{Q}_j$ out of the total objects in the same scene we can match in modality $\mathcal{Q}_k$. This combined measure shows instance matching failure within a scene.

We further evaluate the challenging task of cross-modal scene retrieval within a database. For example, given a
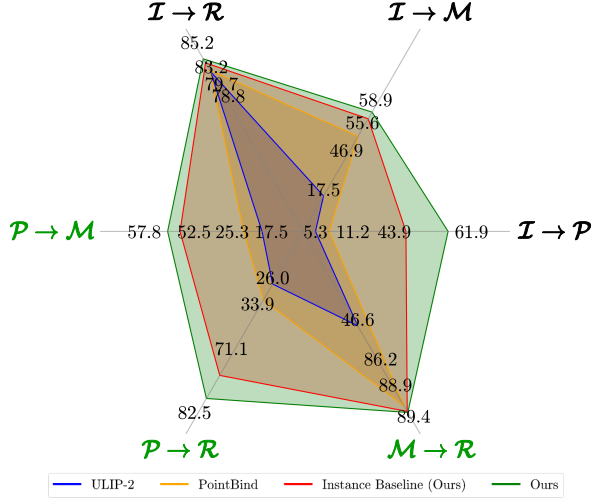
(a) **Instance Matching Recall on ScanNet**

Scene-level Recall ↑	Scannet [11]			3RScan [38]		
	R@25%	R@50%	R@75%	R@25%	R@50%	R@75%
$\mathcal{I} \to \mathcal{P}$						
ULIP-2 [43]	1.28	0.64	0.24	1.91	0.40	0.28
PointBind [18]	6.73	0.96	0.32	3.18	0.64	0.01
Inst. Baseline (Ours)	88.46	37.82	1.92	93.63	35.03	3.82
Ours	**98.08**	**76.92**	**23.40**	**99.36**	**79.62**	**22.93**
$\mathcal{I} \to \mathcal{R}$						
ULIP-2 [43]	98.12	96.21	60.34	98.66	85.91	36.91
PointBind [18]	98.22	95.17	62.07	**100**	87.25	41.61
Inst. Baseline (Ours)	99.31	97.59	71.13	**100**	92.62	55.03
Ours	**99.66**	**98.28**	**76.29**	**100**	**97.32**	**67.79**
$\mathcal{P} \to \mathcal{R}$						
ULIP-2 [43]	37.24	16.90	8.62	16.78	6.04	1.34
PointBind [18]	54.83	27.93	11.72	21.48	6.04	2.01
Inst. Baseline (Ours)	98.63	83.85	46.74	92.62	60.40	20.81
Ours	**99.31**	**96.56**	**70.10**	**100**	**89.26**	**50.34**

(b) **Scene-Level Matching Recall on ScanNet and 3RScan**

Figure 4. **Cross-Modal Instance Retrieval on ScanNet and 3RScan.** (a) Even though CrossOver does not explicitly train all modality combinations, it achieves emergent behavior within the embedding space. The same applies to our Instance Baseline (Ours). CrossOver performs better than our self-baseline since it incorporates more scene context in the fusion of modalities. (b) Our method outperforms all baselines in all datasets, showcasing the robustness of learned cross-modal interactions.
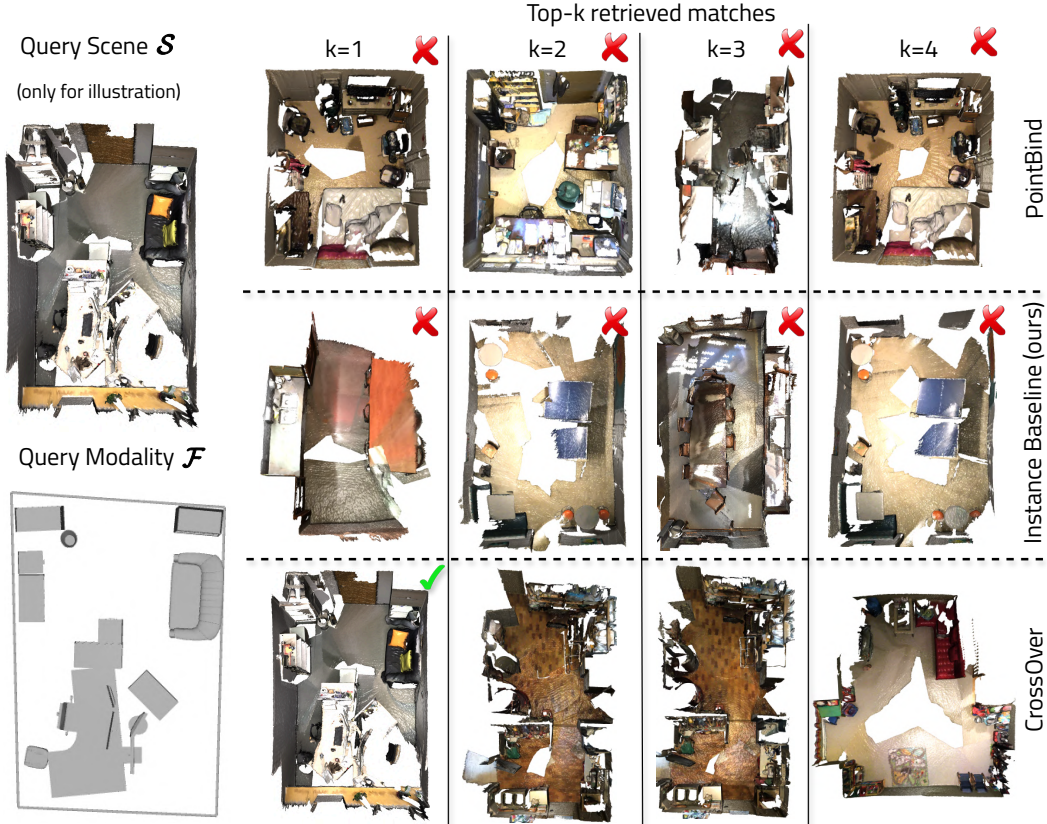

Figure 5. **Cross-Modal Scene Retrieval Qualitative Results on ScanNet.** Given a scene in query modality $\mathcal{F}$, we aim to retrieve the same scene in target modality $\mathcal{P}$. While PointBind and the Instance Baseline do not retrieve the correct scene within the top-4 matches, CrossOver identifies it as the top-1 match. Notably, temporal scenes appear close together in CrossOver's embedding space (*e.g.*, $k = 2$, $k = 3$), with retrieved scenes featuring similar object layouts to the query scene, such as the red couch in $k = 4$.

Scene-level Recall ↑			
Method	**R@25%**	**R@50%**	**R@75%**
same-modal ($\mathcal{P} \to \mathcal{P}$)			
MendNet [15]	80.68	64.77	37.50
VN-DGCNN$_{cls}$ [13]	72.32	53.41	29.55
VN-ONet$_{recon}$ [13]	86.36	71.59	44.32
LivingScenes [47]	87.50	78.41	50.00
Ours	**92.31**	**84.62**	**57.69**
cross-modal (ours)			
$\mathcal{I} \to \mathcal{P}$	89.74	73.08	42.31
$\mathcal{I} \to \mathcal{R}$	62.33	38.96	18.18
$\mathcal{P} \to \mathcal{R}$	68.83	40.26	22.08

Table 1. **Temporal Instance Matching on *3RScan* [38].** Our method exhibits better performance in the same-modal task compared to baselines, despite not being specifically trained on this. It also performs well on cross-modal tasks. Lower performance when $\mathcal{R}$ is involved is expected, as descriptions are contextualized within the scene's layout and may lose validity if objects rearrange.

query point cloud of a scene, we aim to retrieve its corresponding 2D floorplan. This analysis includes multiple levels: (i) *scene matching recall*, or the model's ability to retrieve the exact scene $\mathcal{S}_i$; (ii) *scene category recall* to test retrieval of a scene from the same category (*e.g.*, retrieving *any* kitchen when given a kitchen query in a multi-category database); (iii) *temporal recall* to evaluate whether the model can recover the same scene captured at a different time, accounting for potential object movement or removal; and (iv) *intra-category recall*, which assesses retrieval of a specific scene within a single-category database (*e.g.*, retrieving a particular kitchen from only kitchen scenes). This last metric uniquely requires a different database.

4.1. Instance Retrieval

Cross-Modal Instance Matching. Our goal is instance matching within the same scene where multiple instances of the same furniture (*e.g.*, *two identical chairs*) are commonly present. We showcase our results on *ScanNet* and *3RScan* datasets in Fig. 4. We compare CrossOver with pretrained multi-modal methods ULIP-2 [43] and PointBind [18] and our instance-level multi-modal encoder to highlight the importance of scene-level understanding in a cross-modal embedding space. As shown in Fig. 4a, our performance on ScanNet is robust across modalities, while baselines exhibit varying results. Current multi-modal methods are large pretrained models with strong text encoders that boost performance for *referral*-based retrieval. While prior work trains on all pairwise modalities, we selectively train only in reference to the image modality ($\mathcal{I}$). Yet, we still achieve robust performance across all modalities, even without direct interactions during training. Emergent interactions are in green. Similar trends appear in Fig. 4b for scene-level matching.
Temporal Instance Matching. Although not part of the learning objective, we evaluate CrossOver's effectiveness

Method	Scene Category Recall ↑			Temporal Recall ↑			Intra-Category Recall ↑		
	top-1	top-5	top-10	top-1	top-5	top-10	top-1	top-3	top-5
$\mathcal{I} \to \mathcal{P}$									
ULIP-2 [43]	7.37	25.96	43.27	0.04	1.00	3.00	16.77	41.53	55.54
PointBind [18]	13.78	24.36	42.95	2.00	5.00	7.00	20.03	41.54	57.01
Inst. Baseline (Ours)	42.95	70.19	81.09	13.00	35.00	60.00	46.37	79.68	88.43
Ours	64.74	89.42	94.23	13.00	41.00	84.00	38.98	73.28	85.00
$\mathcal{I} \to \mathcal{R}$									
ULIP-2 [43]	41.92	57.73	61.86	1.00	2.00	8.00	19.48	42.18	56.69
PointBind [18]	49.48	70.45	80.07	2.00	6.00	12.00	19.19	41.54	55.85
Inst. Baseline (Ours)	49.14	71.48	80.07	8.00	28.00	46.00	28.00	62.33	72.62
Ours	57.39	82.82	87.63	3.00	25.00	51.00	29.04	57.85	70.75
$\mathcal{P} \to \mathcal{R}$									
ULIP-2 [43]	11.34	15.12	23.27	1.00	2.00	4.00	18.12	41.15	54.93
PointBind [18]	18.21	26.46	31.96	1.00	2.00	6.00	18.25	40.05	54.84
Inst. Baseline (Ours)	28.87	50.86	66.67	5.00	13.00	23.00	29.41	50.84	65.65
Ours	57.73	79.04	85.57	5.00	20.00	46.00	26.79	56.67	68.63
$\mathcal{I} \to \mathcal{F}$									
ULIP-2 [43]	38.46	55.77	64.42	1.00	2.00	10.00	18.48	39.09	55.96
PointBind [18]	35.58	62.82	72.76	1.00	11.00	21.00	20.03	43.08	58.62
Ours	58.01	81.09	89.10	8.00	32.00	61.00	28.57	55.67	71.77
$\mathcal{P} \to \mathcal{F}$									
ULIP-2 [43]	13.14	26.28	33.65	1.00	1.00	6.00	17.46	38.74	53.99
PointBind [18]	14.10	48.72	59.62	0.50	5.00	7.00	23.17	39.23	57.08
Ours	55.77	78.53	86.54	10.00	30.00	57.00	31.34	63.42	74.15
$\mathcal{R} \to \mathcal{F}$									
ULIP-2 [43]	8.25	29.21	40.21	1.00	2.00	5.00	18.24	41.80	55.35
PointBind [18]	14.43	27.15	48.45	1.00	5.00	8.00	13.64	38.32	54.20
Ours	54.64	74.91	80.41	6.00	17.00	35.00	23.00	51.37	66.84

Table 2. **Cross-Modal Scene Retrieval on *ScanNet*.** We consistently outperform state-of-the-art methods and our self-baseline in most cases. The latter performs better in certain modality pairs on *intra-category*, with the biggest gap observed in $\mathcal{I} \to \mathcal{R}$; this can be attributed to our less powerful text encoder.

Method	Scene Matching Recall ↑				Temporal Recall ↑		
	top-1	top-5	top-10	top-20	top-1	top-5	top-10
$\mathcal{I} \to \mathcal{P}$							
ULIP-2 [43]	1.27	5.10	7.01	12.74	0.04	4.26	12.77
PointBind [18]	1.27	4.46	9.55	17.20	2.13	4.26	8.51
Inst. Baseline (Ours)	8.92	30.57	43.31	64.33	0.04	19.15	42.55
Ours	14.01	49.04	66.88	83.44	12.77	36.17	70.21
$\mathcal{I} \to \mathcal{R}$							
ULIP-2 [43]	2.01	4.70	7.38	14.77	2.13	6.38	12.77
PointBind [18]	1.34	4.77	6.71	13.42	2.13	6.38	14.89
Inst. Baseline (Ours)	8.72	40.94	57.05	69.80	6.38	38.30	63.83
Ours	6.04	26.85	42.28	62.42	2.13	34.04	63.83
$\mathcal{P} \to \mathcal{R}$							
ULIP-2 [43]	0.67	3.36	6.71	12.75	2.13	6.38	6.38
PointBind [18]	0.67	3.36	6.71	13.42	2.13	6.38	6.38
Inst. Baseline (Ours)	0.76	14.09	24.83	36.24	0.04	14.89	27.66
Ours	6.71	19.46	32.31	51.01	8.51	27.66	51.06

Table 3. **Cross-Modal Scene Retrieval on *3RScan*.** Similar performance to the ScanNet results in Fig. 6 is observed.

on temporal point cloud-based instance retrieval (same-modal) using scans acquired at different time intervals, with scene changes like object displacement and rearrangement. Tab. 1 shows a comparison on the *3RScan* dataset, highlighting our method's superior performance. This is a large gain, lying in the strong representational power of our multi-modal embedding space, which allows the encoder to efficiently extract each instance's spatial and geometric features in dynamic scenes. Moreover, our method, while primarily evaluated in the same-modal setting, also demonstrates superior performance in the cross-modal scenario, shown in the second half of Tab. 1, further underlining the importance of scene-level multi-modal alignment to handle temporal variations in indoor scene understanding.
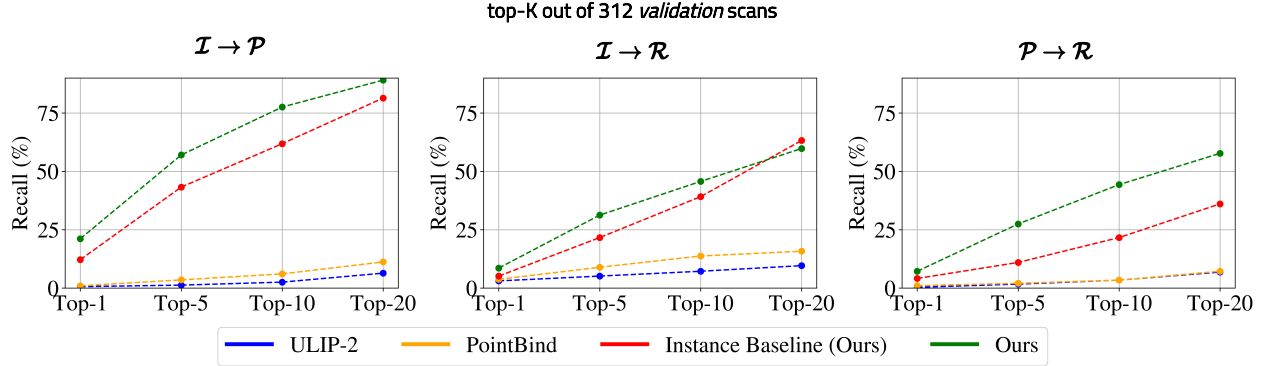
Figure 6. **Cross-Modal Scene Retrieval on *ScanNet* (Scene Matching Recall).** Plots show the top 1, 5, 10, 20 scene matching recall of different methods on three modality pairs: $\mathcal{I} \rightarrow \mathcal{P}, \mathcal{I} \rightarrow \mathcal{R}, \mathcal{P} \rightarrow \mathcal{R}$. Ours and Instance Baseline have not been explicitly trained on $\mathcal{P} \rightarrow \mathcal{R}$. Results are computed on 306 scenes and showcase the superior performance of our approach. Once again, the difference between Ours and our self-baseline is attributed to the enhanced cross-modal scene-level interactions achieved with the unified encoders.

4.2. Cross-Modal Scene Retrieval

We compare our cross-modal scene retrieval results with [18, 43] and our instance-level baseline. Since prior work does not address this task, we adapt their methods by averaging object embeddings per modality to create scene representations, treating our baseline similarly. Unlike CrossOver, these methods rely on semantic instance segmentation. *Scene matching recall* results on *ScanNet* (Fig. 6) show that our unified encoders, not relying on semantics, consistently outperform prior methods in all pairwise modalities and surpass our baseline. Detailed results on *ScanNet* and *3RScan* are in Tabs. 2 and 3. Our method achieves overall scene understanding, even with small-scale object reconfigurations, as shown by its high temporal recall. The lower performance of pretrained methods may stem from training biases that limit their robustness with real-world data, such as incomplete point clouds and blurry images. Qualitative results are in Fig. 5.

4.3. Missing Modalities

To demonstrate CrossOver's ability to capture emergent modality behavior with non-overlapping training data points, we train CrossOver using different data repositories for each modality pair. Specifically, we use the *ScanNet* dataset and split the image repository into two chunks of varying sizes. Training on image-point cloud ($\mathcal{I} \rightarrow \mathcal{P}$) and image-mesh ($\mathcal{I} \rightarrow \mathcal{M}$) using each chunk respectively, we expect to see an emergent behavior between point cloud and mesh ($\mathcal{P} \rightarrow \mathcal{M}$). The results (Tab. 4) include top-1 and top-3 instance matching recall, as well as *same* and *diff* recall for evaluating intra- (e.g., identical chairs) and inter- (e.g., a chair and a table) object category performance within a scene. Although partial data availability decreases recall, our $\mathcal{P} \rightarrow \mathcal{M}$ matching only decreases by 3% even when using 25% $\mathcal{I} \rightarrow \mathcal{P}$. This scenario is common in real-world applications, where certain modalities might be scarce.

Available Data		Instance Matching Recall ↑			
$\mathcal{I} \rightarrow \mathcal{P}$ (%)	$\mathcal{I} \rightarrow \mathcal{M}$ (%)	same	diff	top-1	top-3
25	75	86.32	**73.38**	55.46	79.73
50	50	**87.46**	70.02	57.49	79.94
75	25	87.35	67.65	54.99	79.45
100	100	87.44	72.46	**59.88**	**80.81**

Table 4. **Ablation on $\mathcal{P} \rightarrow \mathcal{M}$ instance matching on *ScanNet* with non-overlapping data per modality pair.** Despite modality pairs not sharing the same image repository, our method retains high performance even when a pair is underrepresented in the data.

5. Conclusion

In summary, this work introduces CrossOver, a framework for flexible, scene-level cross-modal alignment without the need for semantic annotations or perfectly aligned data. CrossOver leverages a unified embedding space centered on image features, allowing it to generalize across unpaired modalities and outperform existing methods in cross-modal scene retrieval and instance matching on real-world datasets. This approach addresses the limitations of traditional multi-modal models and holds promise for practical applications in areas like robotics, AR/VR, and construction monitoring. Although CrossOver excels in cross-modal instance matching, its scene retrieval generalizability could benefit from training on diverse indoor and outdoor datasets. CrossOver assumes a base modality per dataset, advancing prior work requiring perfect modality alignment. Further relaxation is a promising direction. Finally, exploring its embedding space for downstream scene understanding remains a key area. Future research can explore how our approach can be applied to dynamic scene reconstruction and real-time navigation, thus leading to interactive and immersive mixed-reality experiences.

6. Acknowledgements

References

[1] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas J. Guibas. ReferIt3D: Neural listeners for fine-grained 3d object identification in real-world scenes. In *16th European Conference on Computer Vision (ECCV)*, 2020. 2

[2] Relja Arandjelović, Petr Gronát, Akihiko Torii, Tomás Pajdla, and Josef Sivic. Netvlad: Cnn architecture for weakly supervised place recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5297–5307, 2015. 2

[3] 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 International Conference on Computer Vision*, pages 5664–5673, 2019. 2

[4] Armen Avetisyan, Manuel Dahnert, Angela Dai, Manolis Savva, Angel X. Chang, and Matthias Niessner. Scan2cad: Learning cad model alignment in rgb-d scans. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[5] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 1, 2

[6] Angel X. Chang, Thomas A. Funkhouser, Leonidas J. Guibas, Pat Hanrahan, Qi-Xing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, L. Yi, and Fisher Yu. Shapenet: An information-rich 3d model repository. *ArXiv*, abs/1512.03012, 2015. 2, 5

[7] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020. 2

[8] Jiaqi Chen, Daniel Barath, Iro Armeni, Marc Pollefeys, and Hermann Blum. "where am i?" scene retrieval with language. *ArXiv*, abs/2404.14565, 2024. 2

[9] Shizhe Chen, Pierre-Louis Guhur, Makarand Tapaswi, Cordelia Schmid, and Ivan Laptev. Language conditioned spatial relation reasoning for 3d object grounding. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 4

[10] Christopher Choy, JunYoung Gwak, and Silvio Savarese. 4d spatio-temporal convnets: Minkowski convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3075–3084, 2019. 5

[11] 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. 5, 6, 14

[12] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers, 2023. 4, 14

[13] 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 *CVPR*, 2021. 7

[14] Mingyu Ding, Zhe Wang, Jiankai Sun, Jianping Shi, and Ping Luo. Camnet: Coarse-to-fine retrieval for camera relocalization. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2871–2880, 2019. 2

[15] Shivam Duggal, Zihao Wang, Wei-Chiu Ma, Sivabalan Manivasagam, Justin Liang, Shenlong Wang, and Raquel Urtasun. Mending neural implicit modeling for 3d vehicle reconstruction in the wild. In *WACV*, 2022. 7

[16] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. Clip-adapter: Better vision-language models with feature adapters. *arXiv preprint arXiv:2110.04544*, 2021. 2

[17] 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 *CVPR*, 2023. 2, 3, 5, 15

[18] 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. 1, 2, 4, 6, 7, 8, 11, 13, 14

[19] Georg Hess, Adam Tonderski, Christoffer Petersson, Kalle Åström, and Lennart Svensson. Lidarclip or: How i learned to talk to point clouds. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2024. 12, 16

[20] Baoxiong Jia, Yixin Chen, Huangyue Yu, Yan Wang, Xuesong Niu, Tengyu Liu, Qing Li, and Siyuan Huang. Sceneverse: Scaling 3d vision-language learning for grounded scene understanding. *arXiv preprint arXiv:2401.09340*, 2024. 3, 4, 5, 14, 15

[21] Ue-Hwan Kim, Jin-Man Park, Taek-Jin Song, and Jong-Hwan Kim. 3-d scene graph: A sparse and semantic representation of physical environments for intelligent agents. *IEEE transactions on cybernetics*, 50(12):4921–4933, 2019. 2

[22] Manuel Kolmet, Qunjie Zhou, Aljosa Osep, and Laura Leal-Taixe. Text2Pos: Text-to-Point-Cloud Cross-Modal Localization . In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6677–6686, Los Alamitos, CA, USA, 2022. IEEE Computer Society. 2

[23] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022. 3, 14

[24] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 14

[25] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004. 5

[26] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 2

[27] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 2

[28] Yiwei Ma, Guohai Xu, Xiaoshuai Sun, Ming Yan, Ji Zhang, and Rongrong Ji. X-CLIP:: End-to-end multi-grained contrastive learning for video-text retrieval. *arXiv preprint arXiv:2207.07285*, 2022. 2

[29] 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 *EuropeanConference on Computer Vision (ECCV)*, 2024. 2, 4, 12, 14, 15, 16

[30] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 2

[31] 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, 2023. 4, 14

[32] 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 *International Conference on Machine Learning*, 2021. 2, 3, 14

[33] Antoni Rosinol, Arjun Gupta, Marcus Abate, Jingnan Shi, and Luca Carlone. 3d dynamic scene graphs: Actionable spatial perception with places, objects, and humans. *arXiv preprint arXiv:2002.06289*, 2020. 2

[34] Sayan Deb Sarkar, Ondrej Miksik, Marc Pollefeys, Daniel Barath, and Iro Armeni. Sgaligner : 3d scene alignment with scene graphs. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2023. 1, 2, 4, 12

[35] Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning feature matching with graph neural networks. In *CVPR*, 2020. 5

[36] Sai Shubodh, Mohammad Omama, Husain Zaidi, Udit Singh Parihar, and Madhava Krishna. Lip-loc: Lidar image pre-training for cross-modal localization. 2024. 12, 16

[37] Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations. *arXiv preprint arXiv:1806.06176*, 2018. 2

[38] Johanna Wald, Armen Avetisyan, Nassir Navab, Federico Tombari, and Matthias Nießner. Rio: 3d object instance relocalization in changing indoor environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7658–7667, 2019. 5, 6, 7, 14

[39] Johanna Wald, Helisa Dhamo, Nassir Navab, and Federico Tombari. Learning 3d semantic scene graphs from 3d indoor reconstructions. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[40] Renjie Wu, Hu Wang, and Hsiang-Ting Chen. A comprehensive survey on deep multimodal learning with missing modality. *arXiv preprint arXiv:2409.07825*, 2024. 2

[41] 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. *arXiv preprint arXiv:2403.19474*, 2024. 1, 2

[42] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning unified representation of language, image and point cloud for 3d understanding. *arXiv preprint arXiv:2212.05171*, 2022. 1, 2

[43] Le Xue, Ning Yu, Shu Zhang, 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, 2023. 1, 2, 4, 6, 7, 8, 11, 13, 14

[44] Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas J. Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Dvt: Denoising vision transformers. *arXiv preprint arXiv:2401.02957*, 2024. 4

[45] 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. *arXiv preprint arXiv:2112.02413*, 2021. 1, 2

[46] Renrui Zhang, Liuhui Wang, Yu Jiao Qiao, Peng Gao, and Hongsheng Li. Learning 3d representations from 2d pre-trained models via image-to-point masked autoencoders. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21769–21780, 2022. 4, 14

[47] Liyuan Zhu, Shengyu Huang, and Iro Armeni Konrad Schindler. Living scenes: Multi-object relocalization and reconstruction in changing 3d environments. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 7

[48] Zhu Ziyu, Ma Xiaojian, Chen Yixin, Deng Zhidong, Huang Siyuan, and Li Qing. 3d-vista: Pre-trained transformer for 3d vision and text alignment. In *ICCV*, 2023. 4, 14