CosCAD: Cross-Modal CAD Model Retrieval and Pose Alignment From a Single Image

Zhikun Wen [1] Honghua Chen[1] Zhe Zhu[1] Zeyong Wei[1] Liangliang Nan[2]
Mingqiang Wei[1(✉)]
[1] Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2] Delft University of Technology, Delft, Netherlands
mqwei@nuaa.edu.cn

Abstract

We introduce CosCAD, a novel framework for CAD model retrieval and pose alignment from a single image. Unlike previous methods that rely solely on image data and are sensitive to occlusion, CosCAD leverages cross-modal contrastive learning to integrate image, CAD model, and text features into a shared representation space. This improves retrieval accuracy, even when visual cues are ambiguous or objects are partially occluded. To enhance retrieval efficiency, we introduce Tri-Indexed Quantized Graph Search, which accelerates CAD retrieval using an optimized indexing structure. For pose alignment, we combine image and geometric features of CAD models to predict object rotation and scale, using an attention-based method to capture spatial correlations within the scene. This improves multi-object location estimation and 9-DoF pose alignment. Experimental results demonstrate that CosCAD outperforms existing methods such as ROCA and SPARC in both CAD model retrieval and pose estimation. Additionally, it achieves more than a sixfold speedup in retrieval for large datasets, underscoring its potential for interactive environments and autonomous systems.

Keywords: ***CAD Model retrieval, CAD Model alignment, Contrastive learning, Attention***

1. Introduction

Advances in 2D perception systems have achieved great success in object recognition, localization, and classification in images [21, 32, 42], driving progress in autonomous vehicles, machine vision, and virtual/augmented reality. Despite these advancements, inferring 3D geometry, structure, and object poses from a single RGB image remains a challenge. Traditional 2D systems rely on partial views and visual observations, which are insufficient for fully capturing the 3D information required for applications involving interaction with the environment.

Recent studies have explored predicting 3D geometry and pose from 2D visual data [11, 15, 19, 45, 46, 8, 51]. For example, Mesh R-CNN [18] introduced a method for 3D object estimation from real-world images by combining 2D object detection with voxel-to-mesh estimation to reconstruct object shapes. However, Mesh R-CNN lacks explicit pose estimation and relies heavily on 2D feature regression, which limits its accuracy. In contrast, methods like Mask2CAD [25, 37], ROCA [20], and SPARC [27] retrieve and align CAD models directly to the input image. Databases such as ShapeNet, 3DFuture, and 3DShapeNet [5, 16, 47] serve as object priors, making retrieval-based methods promising for inferring 3D information from images.

However, retrieval-based methods have limitations: (1) These methods rely solely on image features, making the retrieval process highly sensitive to occlusion, where portions of objects in the image may be hidden. This significantly reduces retrieval precision. (2) The retrieval process usually involves matching the image feature with each feature of the CAD model one by one, requiring an exhaustive search within the database. This significantly reduces retrieval efficiency. (3) These methods focus on the pose estimation of individual objects, neglecting relationships between objects in the scene. However, understanding these relationships can provide valuable contextual constraints and help resolve ambiguities in object poses, as objects often appear in similar scene layouts.

To address these limitations, we propose CosCAD, a novel framework for retrieving and aligning CAD models from a single 2D image. For CAD model retrieval, we adopt cross-modal contrastive learning to unify representations of image features, CAD models, and text (e.g., category labels) into a shared representation space. This unified representation enables the use of combined image and text inputs to enhance retrieval, with text providing cru-

cial semantic information to complement CAD model retrieval, especially when visual cues are unclear or incomplete. To speed up retrieval and avoid exhaustive search through the CAD database, we propose the Tri-Indexed Quantized Graph Search (TQGS) method, which constructs an efficient index structure for the feature vectors of CAD models. For CAD model alignment, we combine image and geometric features to predict the rotation and scale of an object. We introduce an attention-based method that explicitly captures positional correlations between multiple objects in the scene. By leveraging these correlations, our approach captures crucial contextual information, enabling more accurate multi-object location predictions.

Our core contributions are summarized as follows:

- Developing a shared cross-modal representation from 3D models, text, and images to enable accurate CAD model retrieval.
- Proposing the Tri-Indexed Quantized Graph Search method that accelerates CAD model retrieval.
- Leveraging an attention mechanism to exploit spatial correlations among objects, which results in accurate location estimation.

2. Related Work

2.1. 2D Object Recognition

Understanding 3D from a single image requires robust 2D recognition capabilities. Thanks to recent breakthroughs in deep learning, an array of methods has emerged that demonstrate exceptional performance in tasks such as image classification [22, 42], 2D object detection [32, 33], and 2D instance segmentation [21, 26]. Recent methods such as YOLOv8 [24] and EfficientDet [44] have significant improved detection speed and accuracy, which are crucial for real-time applications. Additionally, advancements like Detection Transformers [4] have introduced end-to-end object detection models that eliminate the need for hand-designed components, further streamlining the object detection pipeline. Our approach leverages these advances in 2D recognition to facilitate comprehensive 3D object reasoning and enhance the accuracy of CAD model alignment and retrieval. Specifically, we employ a Mask R-CNN [21] recognition backbone for image feature extraction, which not only detects but also segments objects within the image. This subsequently aids in 3D CAD alignment and retrieval. Additionally, Mask R-CNN is known for its efficiency in handling both detection and segmentation tasks in a single forward pass. Compared to other methods, Mask R-CNN integrates region proposal and mask prediction more effectively, reducing computational overhead. This efficiency enables real-time retrieval and alignment of CAD models from a single image.

2.2. Multimodal Representation Learning

Multimodal representation learning has gained significant attention recently because of its ability to integrate various types of data, such as images and text, which improves model understanding and prediction capabilities. Unlike unimodal approaches that rely on a single data type, such as either image or text, multimodal methods explore the complex interactions between different data types, resulting in richer and more robust representations. Notably, some studies focus on image-text pairs, using transformer-based architectures to capture the interactions between image and text [43, 35, 31, 29, 28, 9]. Despite their effectiveness, these models often require substantial computational resources for training.

Alternatively, models like CLIP [40] and SLIP [38] have streamlined the learning process by separately encoding images and text and then aligning their representations across modalities. This alignment strategy, coupled with the ability to leverage large-scale noisy datasets, has facilitated efficient training and robust zero-shot learning. These capabilities have enabled new multimodal tasks, such as text-guided image editing, object detection without predefined categories, and grounding language in visual contexts [30]. Additionally, incorporating multimodal information beyond text and image, such as audio or video, has been shown to further benefit 3D scene understanding. A recent advancement, PointCLIP [49], applies these multimodal principles to 3D data by converting point clouds into depth maps, subsequently utilizing CLIP for zero-shot 3D classification. While PointCLIP aligns 3D data with 2D images and text, our method creates a unified representation that integrates images, text, and CAD models, enhancing CAD model understanding in both depth and accuracy.

2.3. CAD Model Retrieval and Alignment

The use of CAD model priors for 3D reconstruction has been a fundamental approach in computer vision for many years [3, 10, 41, 6, 7]. With the introduction of large-scale 3D shape datasets [5], numerous techniques have emerged that emphasize CAD model retrieval and alignment through analysis-by-synthesis methods [48]. Advances in deep learning, particularly with Convolutional Neural Networks (CNNs) and Graph Neural Networks (GNNs), have significantly enhanced the precision and efficiency of these approaches. Points2Objects [14] advances beyond basic 2D object detection [13] by directly predicting 9-DoF alignments and framing object retrieval as a classification task, primarily demonstrating effectiveness on synthetic datasets. In contrast, Mask2CAD [25] provides an efficient method to simultaneously retrieve and align 3D CAD models to detected objects in an image, leveraging a state-of-the-art 2D recognition backbone [26, 33]. This innovative use of CAD representations for lightweight, object-based reconstruction

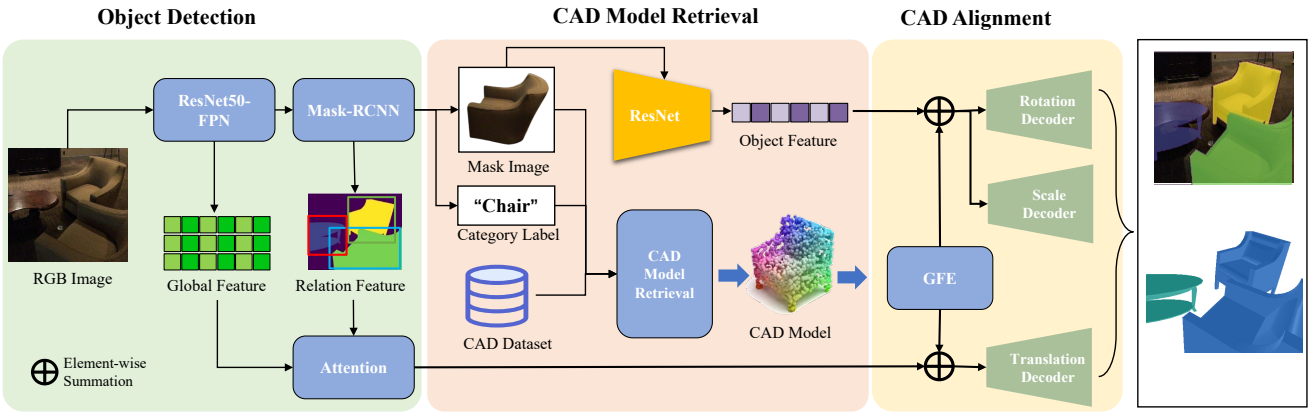

Figure 1. Pipline of the proposed CosCAD. Starting with an input RGB image, we perform 2D instance segmentation and assign category labels to each object. For each detected object, we unify the representations of text, images, and CAD models to retrieve geometrically similar CAD models. In the CAD model alignment stage, we integrate image features with the geometric features of CAD models to predict the rotation and scale of each object and estimate their locations using an attention-based method. This enables geometrically informed CAD retrieval and robust alignment with the image.

and perception has led to the development of patch-based methods for improved CAD retrieval and its extension to video inputs in Vid2CAD [37], which offers comprehensive 9-DoF alignment for each object.

Building on these advancements, ROCA [20] proposed a CAD retrieval and alignment method that incorporates a differentiable 9-DoF pose optimization, representing a significant step forward and offering a more robust and geometry-aware solution for end-to-end CAD alignment. Unlike traditional methods that rely on straightforward regression models, ROCA introduces a novel approach with differentiable alignment through dense geometric correspondences, allowing for nearest neighbor retrieval of objects from a CAD model database. Further extending this work, SPARC [27] adopts a render-and-compare strategy using efficient transformer architectures to achieve more accurate and robust 9-DoF pose estimation. This approach utilizes a novel iterative process, substantially enhancing alignment accuracy on real-world datasets.

Our approach aims to retrieve and align CAD models for reconstructing objects in RGB images. Unlike existing techniques that typically focus on a single data modality for CAD model prediction, our method employs a cross-modal strategy by integrating visual and textual data. This integration enhances the accuracy and robustness of CAD model retrieval from extensive databases. For CAD alignment, instead of directly regressing the object pose, we combine geometric features from the retrieved CAD model with image features from each object mask to predict object rotation and scale. Additionally, we propose an attention-based location estimation method that leverages positional correlations among objects to accurately predict their locations.

3. Method

3.1. Overview

Given an RGB image I with camera intrinsics π and a database of 3D CAD models S, our goal is to represent each object in the image with a corresponding CAD model, aligned with 9-DoF alignment to the image, to provide a comprehensive and lightweight geometric scene reconstruction. Figure 1 shows an overview of our approach.

We first detect and segment objects in the 2D image using a Mask R-CNN backbone [21], estimate category labels and extract global features via a multi-scale Feature Pyramid Network (FPN) [32]. Next, we unify the representations of text (category labels), masked images, and CAD models to retrieve geometrically similar CAD models for the detected objects. To accelerate retrieval process and avoid exhaustive searches through the CAD database, we construct an efficient indexing structure for the feature vectors of CAD models. Then, our method predicts each object's rotation and scale by integrating image features with the geometric features of CAD models. The translation is estimated using an attention-based location estimation module that leverages spatial relationships between objects. Our network is trained progressively. It starts with CAD model predictions using object masks and category labels. Next, it predicts object poses through a dedicated network supervised by CAD models aligned with RGB images.

3.2. Object Detection

We employ a ResNet-50 architecture [22] combined with a Feature Pyramid Network (FPN) [32] as the backbone. This backbone generates a feature map F, which is then used for instance segmentation through Mask R-

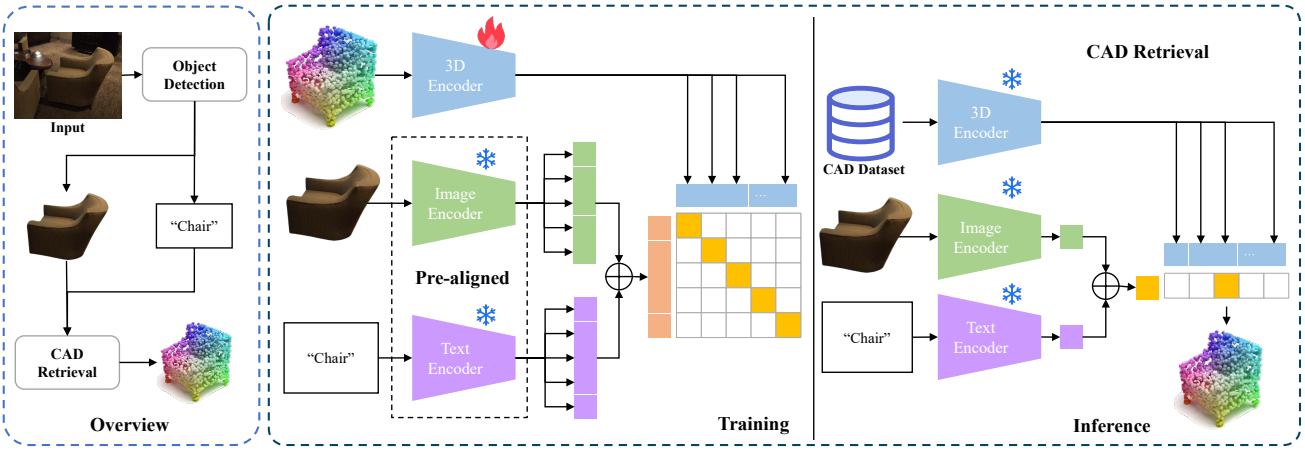

Figure 2. Overview of CAD Model Retrieval. In the training phase, given triplets of CAD models, text, and images, we train a 3D encoder to align 3D features with the image features obtained from SLIP's pre-trained image encoder and the text features derived from its pre-trained text encoder, using a contrastive loss. In the inference phase, the image mask is encoded into image features, while the category label is transformed into text features. By combining these two types of features, we form the query feature, which is then used to retrieve the most similar CAD models from the CAD feature database through feature comparison.

CNN. Mask R-CNN first generates region proposals using a Region Proposal Network (RPN) on the feature map F. These proposals are refined through Region of Interest (ROI) Alignment, ensuring accurate feature alignment with the proposed regions. Each ROI is then processed by a series of fully connected layers to predict the object category and bounding box, producing a set of instance masks m_i along with their corresponding category labels t_i. The object recognition process identifies objects within the image, which provide segmentation masks and category labels essential for the subsequent stages.

3.3. Cross-Modal CAD Model Retrieval

For each detected object, we retrieve its corresponding CAD model. To align 3D representations with holistic image-text pairs, we employ cross-modal contrastive learning to unify the representations of image features, CAD models, and text (e.g., category labels), into a shared representation space. As shown in Figure 2, the 3D point cloud sampled from the surface of the CAD model is used as input for the 3D encoder, the masked image is fed into the 2D encoder, and the category label is processed by the text encoder. Using these triplets, we perform pre-training to align the representations of all three modalities within a shared feature space. Specifically, we leverage the pre-trained vision-language model SLIP [38] and freeze its parameters during pre-training. We then train a 3D encoder by aligning its output features with those of SLIP's image encoder $E_I(\cdot)$ and text encoder $E_T(\cdot)$. This approach ensures that the rich semantics captured by SLIP's encoders are transferred to enhance 3D understanding. The resulting unified feature space enables various cross-modal applications across the three modalities, potentially improving the CAD recognition performance of the underlying 3D backbone encoder $E_P(\cdot)$.

Cross-Modal Contrastive Learning. As illustrated in Figure 2, during tri-modal pre-training, given a CAD model C, a masked instance M, and the corresponding category label T, we uniformly sample N_P points to generate the corresponding point cloud P. During training, we apply data augmentations to P, including random point dropout, scaling of the point cloud, point displacement, and rotational perturbations. These augmentations improve the model's robustness and generalization. Next, we extract the image feature $f^I = E_I(M)$ and text feature $f^T = E_T(T)$. Our objective is to train the 3D encoder E_P so that its 3D feature $f^P = E_P(P)$ aligns with the corresponding image and text feature space. To achieve this, we utilize a contrastive loss to minimize the distance between the 3D feature and the corresponding image and text feature.

To strengthen the association with 3D features, we weight the text and image features, enhancing their combined utilization. Here, we form a query feature f^Q by combining the image and text features as follows:

$$f^Q = \alpha f^I + \beta f^T, \tag{1}$$

where α and β are hyperparameters. Then, the contrastive loss is defined as:

$$\mathcal{L}_{con} = -\frac{1}{2}\sum_i \log \frac{\exp(f_i^P f_i^Q/\tau)}{\sum_j \exp(f_i^P f_j^Q/\tau)} - \frac{1}{2}\sum_i \log \frac{\exp(f_i^P f_i^Q/\tau)}{\sum_j \exp(f_j^P f_i^Q/\tau)}, \tag{2}$$

where i and j represent the sampling indices, and τ is a learnable temperature parameter. The first term ensures the

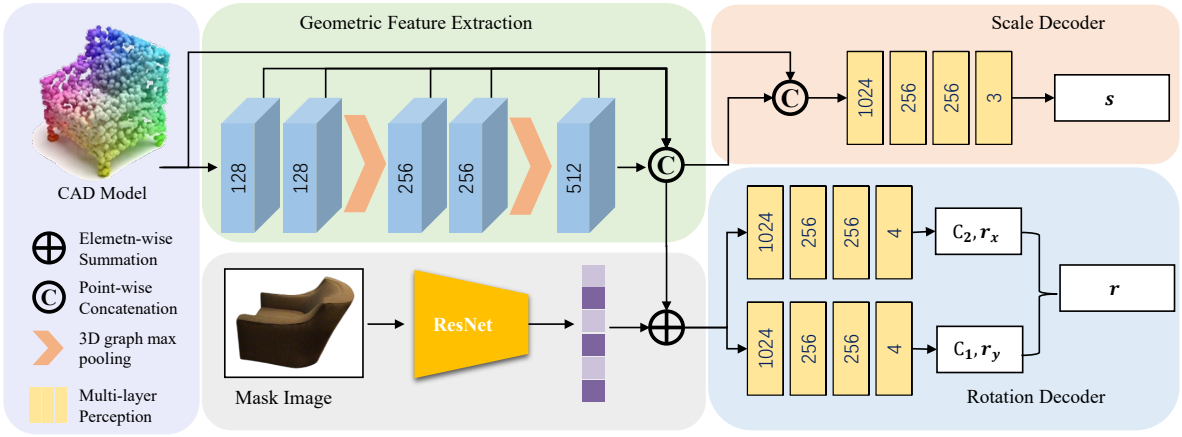

Figure 3. Rotation and Scale Estimation. The Geometric Feature Extraction module extracts 3D geometric features from the input CAD model, retrieved using the mask and category label. For mask feature extraction, we apply the same method used in the object detection phase, with ResNet as the 2D backbone. We calculate the geometric features for each predicted CAD model and the image features for each mask. The geometric and image features are combined by element-wise addition, and MLP layers are used to regress plane normals and their confidence. The rotation is determined based on the regressed plane normals and the scale is regressed using a four-layer MLP.

alignment of the 3D feature with its corresponding query feature while minimizing its similarity to query features from other samples. Similarly, the second term ensures the alignment of the query feature with its corresponding 3D feature, while reducing its similarity to 3D features from other samples.

Tri-Indexed Quantized Graph Search. Retrieving CAD models typically involves matching the image feature vector to each CAD model feature in an exhaustive database search. In a database with 3,000 entries, each with a 512-dimensional feature vector, the worst-case retrieval time can be 1.2 seconds. To accelerate model retrieval, we propose a Tri-Indexed Quantized Graph Search. We first index each CAD model by encoding its 3D point cloud P into a feature vector $f^P = E_P(P)$ using a 3D encoder. We then apply Locality Sensitive Hashing (LSH) [17] to the feature vector, grouping similar features into buckets. For a given feature vector f^P, the LSH hash function $h_i(f^P)$ is defined as:

$$h_i(f^P) = \left\lfloor \frac{a_i^T f^P + b_i}{w} \right\rfloor \tag{3}$$

where a_i is a randomly generated vector, b_i is an offset, and w is the bin width. This method efficiently partitions the vector space, narrowing the search to a smaller subset of CAD models that fall within the same bucket as the query.

After identifying the relevant LSH bucket, we organize its contents using the Hierarchical Navigable Small World Graph (HNSW) [36]. In HNSW, nodes represent CAD models, and edges connect their nearest neighbors, forming a multi-layered graph. The upper layers are sparse, enabling fast pruning of irrelevant nodes, while the lower layers become increasingly dense, allowing for more accurate fine-grained search. The layers are constructed probabilistically, with each layer connecting nodes based on an Euclidean distance measure. During the search, the query feature f^Q begins at the top layer and navigates through the graph, refining its search by identifying closer nodes based on a distance measure, typically Euclidean distance $d(f^P, f^Q) = ||f^P - f^Q||_2$, until it reaches the bottom layer, where the nearest neighbors are located. To further improve retrieval efficiency and reduce storage, we apply Product Quantization (PQ) [23] to the feature vectors. The feature vector f^P is divided into M sub-vectors, defined as:

$$f^P = [f_1^P, f_2^P, ..., f_M^P] \tag{4}$$

Each sub-vector f_i^P is quantized by identifying the closest centroid from a codebook C_i, which is built using centroids learned from the training data with a clustering algorithm like k-means. These centroids represent common patterns within the sub-vector space. The quantization function is defined as follows:

$$q(f_i^P) = \arg\min_{c_{ij} \in C_i} \|f_i^P - c_{ij}\|_2^2 \tag{5}$$

This produces a compact representation of the feature vector, allowing for efficient distance computations during the matching phase.

During inference, we first hash the query feature f^Q into the corresponding LSH bucket using the same LSH function. The query then navigates the HNSW structure to identify candidate CAD models. Finally, we compute the PQ-based distance between the query feature f^Q and the candidate feature vectors f^P as

$$d(f^P, f^Q) \approx \sum_{i=1}^{M} \|q(f_i^P) - q(f_i^Q)\|_2^2 \tag{6}$$

Finally, the CAD model with the smallest distance is returned as the retrieval result, ensuring efficient and accurate CAD model retrieval.

3.4. Multi-Object CAD Model Alignment

During CAD model alignment, we combine image features with the geometric features of CAD models to predict the rotation and scale of each object. We introduce an attention-based method that explicitly captures positional correlations among multiple objects. These correlations provide crucial contextual information that enhances multi-object location predictions. For each object, we estimate its 9-DoF alignment with the image, including translation $t \in \mathbb{R}^3$, scale $s \in \mathbb{R}^3$, and rotation $r \in \mathbb{R}^3$. Figure 3 provides an overview of the module.

Geometric Feature Extraction. Since geometric features are essential for pose estimation across different shapes, we use a Geometric Feature Extraction module (GFE) to extract features from the input CAD model C, retrieved based on the mask and category label. Our GFE employs a hybrid feature extraction layer from [50], which extends 3D graph convolution to extract hybrid latent features from point cloud data. The core component is a deformable kernel that generalizes the convolution kernel from 2D image processing to handle unstructured point cloud data. In particular, a GFE kernel K^S is defined as

$$K^S = \{(k_C, w_C), (k_1, w_1), ..., (k_S, w_S)\}, \tag{7}$$

where S represents the total number of support vectors, $k_C = [0, 0, 0]^T$ is the central kernel point, and $\{k_s \in \mathbb{R}^3\}_{s=1}^S$ denotes the support kernel vectors in 3D space. Each kernel vector is associated with a corresponding weight w. GFE applies convolution over the receptive field $R^M(p_i)$, which includes the target point and its neighboring points, alongside their respective features f as:

$$R^M(p_i) = \{(p_i, f_i), (p_m, f_m) | p_m \in N^M(p_i)\}, \tag{8}$$

where $N^M(p_i)$ represents the set of M nearest neighbors of the point p_i. Specifically, the receptive field is determined using a feature distance metric $dist_f(p_i, p_j) = ||f_i - f_j||$. The key advantage of such design is to extend beyond local regions, enabling the inclusion of distant points that share similar features.

Rotation and Scale Estimation. Once equipped with all required features, we then combine them by element-wise addition to estimate the scale and rotation parameters. To accurately represent a 3D object's rotation, it is essential to describe its orientation in space using three mutually perpendicular axes, which define the object's local coordinate system. Each axis represents a direction in 3D space, and together, they capture the object's complete rotational state. Instead of predicting all three axes separately, we regress two orthogonal plane normals, as the third axis can be uniquely determined from their cross product. This approach guarantees that the axes remain orthogonal and form a valid rotation matrix.

Specifically, we use MLP layers to regress two planes' normal r_y and r_x along with their corresponding confidences c_y and c_x. To ensure the plane normals remain perpendicular, we minimize the following cost function to calibrate them into $r_{y'}$ and $r_{x'}$.

$$\begin{aligned} \theta_1^*, \theta_2^* &= \arg\min c_y\theta_1^2 + c_x\theta_2^2 \\ s.t. \quad & \theta_1 + \theta_2 + \pi/2 = \theta, \end{aligned} \tag{9}$$

Let θ represent the angle between r_x and r_y, we then compute:

$$\begin{cases} \theta_1^* = \frac{c_x}{c_x + c_y}(\theta - \pi/2) \\ \theta_2^* = \frac{c_y}{c_x + c_y}(\theta - \pi/2) \end{cases} \tag{10}$$

The calibrated plane normals r'_y and r'_x are derived from $(\theta_1^, \theta_2^)$ using the Rodrigues Rotation Formula. Based on these normals, the rotation matrix is computed as $R = [r'_x, r'_y, r'_x \times r'_y]$. The scale s is predicted through a four-layer MLP, and both the scale and rotation are optimized with an L_1 loss function, defined as:

$$\mathcal{L}_{scale} = \|s - s^{gt}\|_1, \tag{11}$$

$$\mathcal{L}_{rot} = \|r - r^{gt}\|_1 \tag{12}$$

Attention-based Translation Estimation. For translation estimation, we argue that each object interacts with its surroundings, so we consider all objects in the scene when predicting its location. This approach allows us to leverage contextual information, improving the accuracy of the predicted object positions. The estimation network is illustrated in Figure 4. For every pair of objects in the input image, we compute a relation feature f^R to describe their geometric similarity [39]. The relation feature f^R is defined as $[\Delta x, \Delta y, \Delta \log w, \Delta \log h]$, where the first two terms represent the normalized differences in the center coordinates between the target and the other object, and the latter two terms capture the scale differences through the log ratios of their widths and heights. The normalized differences in center coordinates are computed as:

$$\Delta x = \frac{x_{1,n} + x_{2,n} - x_{1,m} - x_{2,m}}{2\,(x_{2,m} - x_{1,m})}, \tag{13}$$

$$\Delta y = \frac{y_{1,n} + y_{2,n} - y_{1,m} - y_{2,m}}{2\,(y_{2,m} - y_{1,m})} \tag{14}$$

The log ratios of width and height are given by:

$$\Delta \log w = \log\left(\frac{x_{2,n} - x_{1,n}}{x_{2,m} - x_{1,m}}\right), \tag{15}$$

$$\Delta \log h = \log\left(\frac{y_{2,n} - y_{1,n}}{y_{2,m} - y_{1,m}}\right) \tag{16}$$

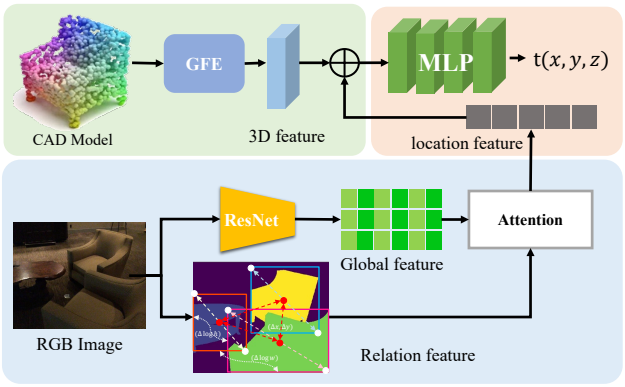

Figure 4. Attention-based Translation Estimation. Image features from object detection are used as global features. For each target object, we compute relational feature to other objects using an object relation module. The attention module processes the relational features and image features to derive the location feature. We then perform element-wise addition of the geometric feature and location features, followed by regressing the translation t using a four-layer MLP.

where $(x_{1,m}, y_{1,m}, x_{2,m}, y_{2,m})$ are the coordinates of the target object's bounding box, and $(x_{1,n}, y_{1,n}, x_{2,n}, y_{2,n})$ are the coordinates of the other object's bounding box.

After obtaining the relation feature among all objects, we employ an attention module to fuse the relation feature f^R and image Global feature f^G, producing the location feature. In this attention process, the relation feature serves as the key and value, while the image feature acts as the query.

Specifically, for each object i, we first extract its initial feature vector f_i^G from the global image features and project it into the query space q_i via a linear projection layer $q_i = W_q f_i^G$, where W_q is a learnable projection matrix. Simultaneously, the relation feature f^R is mapped into the key space k_j and value space v_j by applying linear projections: $k_j = W_k f_j^R$ and $v_j = W_v f_j^G$, where W_k and W_v are also learnable projection matrices. The attention score is computed by the dot product of q_i and k_j, followed by normalization, which serves as a weighted coefficient applied to the value vectors v_j, generating the updated local feature f_i^L:

$$\text{Attention}(q_i, k_j) = \frac{\exp(q_i \cdot k_j / \sqrt{d})}{\sum_{j=1}^{N} \exp(q_i \cdot k_j / \sqrt{d})}, \tag{17}$$

$$f_i^L = \sum_{j=1}^{N} \text{Attention}(q_i, k_j) v_j, \tag{18}$$

where d is the dimension of the feature vectors and N is the total number of objects in the scene. This attention mechanism explicitly models the spatial and geometric relationships between objects, ensuring that the local feature f_i^L incorporates both the target's own information and its interactions with other objects. After obtaining the updated local feature f_i^L, it is element-wise added with the corresponding geometric feature f_i^P of the target object to generate the translation feature:

$$f_i^{\text{trans}} = \frac{f_i^L + f_i^P}{2}. \tag{19}$$

Finally, the translation vector t is predicted using a four-layer MLP. The translation loss is optimized with the following objective function:

$$\mathcal{L}_{trans} = \|t - t^{gt}\|_2 \tag{20}$$

This design allows the model to fully leverage both the spatial geometry and global semantic information, ensuring precise alignment between the CAD model and the target object in the image.

The final alignment loss is defined as:

$$\mathcal{L}_{alignment} = \omega_{trans}\mathcal{L}_{trans} + \omega_{scale}\mathcal{L}_{scale} + \omega_{rot}\mathcal{L}_{rot}, \tag{21}$$

where ω_{trans}, ω_{scale}, and ω_{rot} are hyper-parameters. For the sake of simplicity and to ensure a balanced contribution of each component to the overall loss, we set these hyper-parameters to equal values: $\omega_{trans} = \omega_{scale} = \omega_{rot} = \frac{1}{3}$. This approach enables end-to-end training for alignment, directly influencing the predicted pose derived from the image and CAD model, resulting in more accurate alignment estimates.

4. Experiment

4.1. Data and Evaluation

We compare our method with Mask2CAD-b5 [25], a 9-DOF architecture based on the original Mask2CAD. Mask2CAD-b5 incorporates object depth and scale prediction to produce 9-DoF alignments. Furthermore, we compare our approach with single-image-based 3D object detectors, including Total3D [39] and MDR [34]. Total3D predicts room layouts and object poses and generates object meshes from a single image. For a fair comparison in pose estimation, we evaluate its Object Detection Network. MDR performs joint 3D object detection and voxel-based coarse-to-fine object reconstruction; we compare against its 3D CenterNet-based detector. Since both methods rely on room layout, we provide ground-truth rotations. All baselines are trained on the ScanNet25k dataset.

Dataset. All our experiments are conducted on the ScanNet25k dataset [12], which is also utilized in ROCA [20]. This dataset contains 20k training images and 5k validation images, with both sets sampled from videos across various scenes. For training data preparation, we project the models from Scan2CAD to corresponding image views to generate

Table 1. Alignment Accuracy on ScanNet in comparison to current methods. Total3D-ODN and MDR-CN are the 3D object detectors of Total3D and MDR-CenterNet, respectively. In both detectors, we provide ground-truth rotations in lieu of layout estimation. Mask2CAD-b5 refers to Mask2CAD that predicts full 9-DoF alignment. Our method, ROCA, and SPARC all predict full 9-DoF alignment. In our ablations, CM denotes Corss-Model CAD model retrieval, GFE denotes the Geometric Feature Extraction module, MF denotes the combination of mask features, and ATE denotes Attention-Based Translation Estimation.

Method	bathtub	bed	bin	bkshlf	cabinet	chair	display	sofa	table	class	instance
Total3D-ODN	10.0	2.9	16.8	2.8	4.2	14.4	13.1	5.3	6.7	8.5	10.4
MDR-CN	5.8	5.7	0.9	9.9	5.4	28.1	11.5	11.5	8.1	9.7	15.3
Mask2CAD-b5	8.3	2.9	25.9	3.8	5.4	30.9	17.3	5.3	7.1	11.9	17.9
ROCA	22.5	10.0	29.3	14.2	15.8	41	**30.4**	15.9	14.6	21.5	27.4
SPARC	**25.8**	**25.7**	24.6	14.2	20.8	**51.5**	17.8	**28.3**	15.4	24.9	31.8
Ours CM	23	16.8	27.4	18.0	29.9	43.2	24.2	17.9	21.7	20.2	28.3
Ours CM+GFE	23.5	15.0	29.0	17.3	32.9	43.8	30.1	19.2	20.1	25.6	32.1
Ours CM+GFE+MF	22.1	17.0	28.9	19.9	32.1	45.5	25.3	18.9	22.7	25.8	32.5
Ours CM+GFE+MF+ATE	24.9	18.9	**31.1**	**20.8**	**33.1**	45.4	27.5	19.9	**24.6**	**27.4**	**33.2**

Table 2. Retrieval-Aware Alignment Accuracy on ScanNet.

Method	bathtub	bed	bin	bkshlf	cabinet	chair	display	sofa	table	class	instance
Mask2CAD-b5	7.5	2.9	23.3	2.8	4.2	23.0	12.0	3.5	6.0	9.5	13.8
ROCA	20.8	10.0	26.7	8.5	11.9	32.1	22.5	14.2	11.8	17.6	21.7
SPARC	22.1	13.3	21.7	8.7	15.4	42.1	13.5	19.5	12.5	18.8	25.3
Ours w/o CM	21.3	12.2	18.8	6.9	16.1	33.6	20.6	15.1	10.2	17.2	21.2
Ours	**22.2**	**17.4**	**29.2**	**18.4**	**29.2**	**42.1**	**25.9**	**17.4**	**21.2**	**24.8**	**30.3**

object detection, segmentation, and depth data. The input images are processed at a resolution of 360×480 pixels.

Alignment Accuracy. To assess the 9-DoF alignment performance, we use an alignment accuracy metric similar to that in ROCA [20] and SPARC [27]. The alignment accuracy is defined as follows: an alignment is considered correct if the object classification is accurate, the translation error is $\leq 20cm$, the rotation error is $\leq 20°$, and the scale ratio is $\leq 20\%$. The scale ratio is computed using the formula $s_{error} = |\sum_{i=x,y,z}(s_i^{pred}/s_i^{gt} - 1)|$, allowing errors in different directions to offset each other, unlike the standard formula $s_{error} = \sum_{i=x,y,z}|(s_i^{pred}/s_i^{gt}) - 1|$.

Retrieval-Aware Alignment Accuracy. To further evaluate our method, we use a retrieval-aware alignment accuracy metric inspired by ROCA. This metric takes into account both the alignment correctness and the accuracy of the required CAD model. All CAD candidates are restricted to those in the ScanNet scene, consistent with previous retrieval approaches [1, 37, 2].

4.2. Results

Tables 1 and 2 present the alignment accuracy and retrieval-aware alignment accuracy, respectively. Our method outperforms SPARC in alignment accuracy by 2.5% and improves class and instance averages by 1.4%. Furthermore, in terms of retrieval-aware alignment accuracy, our method demonstrates a significant improvement over SPARC, with gains of 6.0% in class accuracy and 5.0% in instance accuracy.

Figure 5 presents a qualitative comparison of CAD retrieval and alignment on ScanNet images. Our method demonstrates more robust and accurate object alignments across various image views and object types. Figure 6 illustrates the performance of our model in a range of challenging real-world scenarios, demonstrating its effectiveness in accurately predicting 3D structures.

4.3. Ablations

Effectiveness of Cross-Modal CAD model Retrieval. To assess the impact of cross-modal CAD model retrieval on the overall performance, we conducted an ablation study by excluding this component from the CosCAD framework. The resulting model, denoted as Ours w/o CM, struggles to retrieve CAD models. As shown in Table 2, our complete model demonstrates a substantial 7.6% improvement in retrieval-aware alignment accuracy compared to the variant without CM. This decline reveals the critical importance of using cross-modal information (text and image) to improve the retrieval process. The unified representation and alignment of features across modalities play a key role in retrieval, leading to improved alignment accuracy.

Effectiveness of rotation and scale estimation. Another key component of our approach is the Geometric Feature Extraction module (GFE), designed to effectively extract 3D geometric features combined with mask features (MF) for rotation and scale estimation. We evaluate its contribution through an ablation study by incrementally adding the GFE components and MF. The variant without the Geomet-

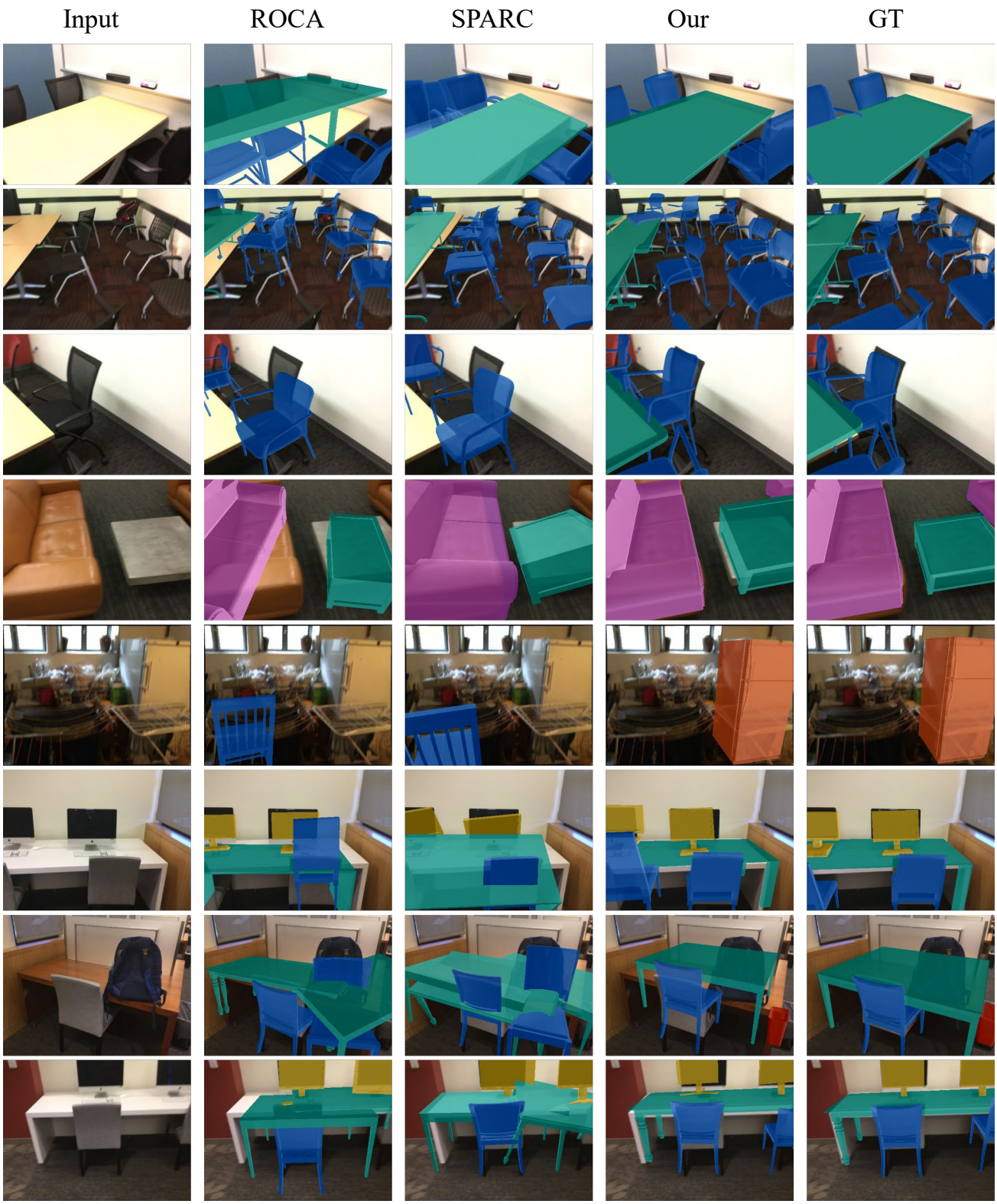

Figure 5. Qualitative comparison on RGB images from ScanNet. We compare CosCAD to ROCA and SPARC in terms of object alignment across various complex scenes. Our approach consistently demonstrates significantly more accurate alignments, especially in challenging environments with diverse objects and occlusions.

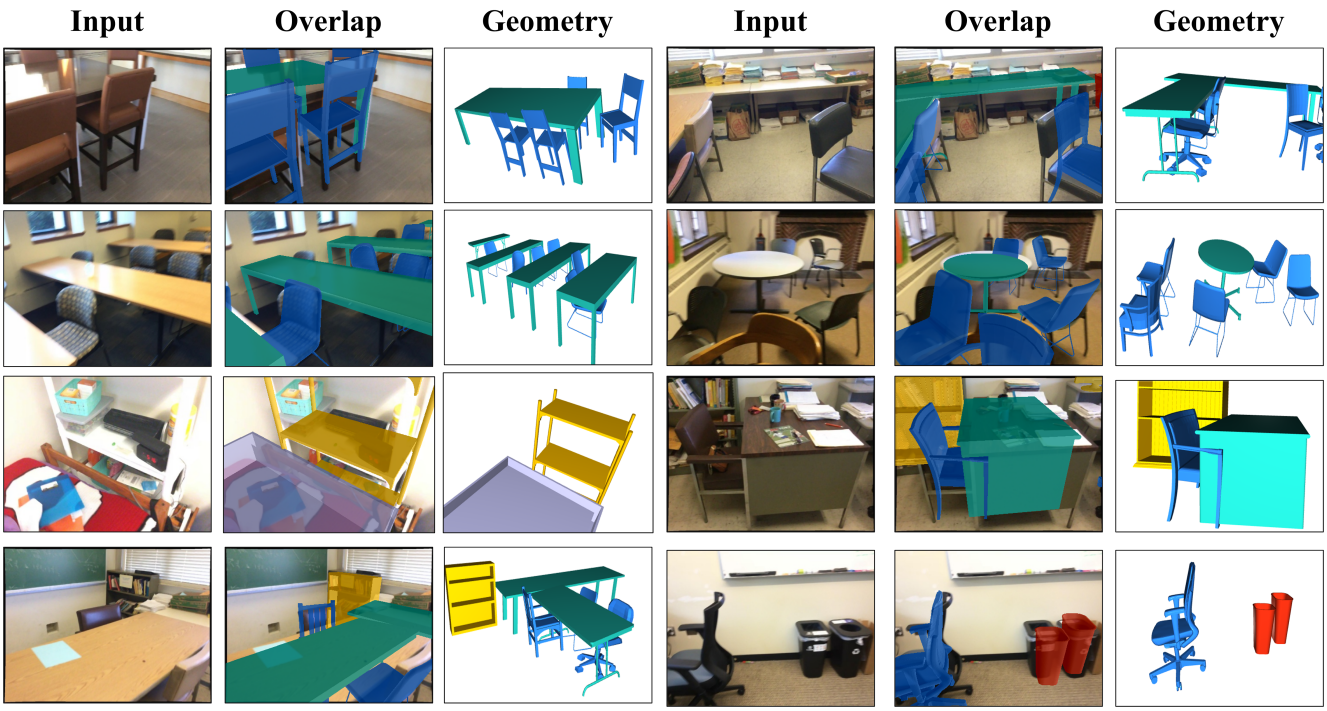

Figure 6. Sampled predictions from our method. The figure illustrates the performance of our model, showing its ability to accurately predict 3D structures from 2D images in a range of challenging real-world scenarios. These examples reveal the robustness of our method in capturing complex geometries, object poses, and spatial relationships across diverse scenes, demonstrating its strong 3D understanding capabilities even in cases with occlusions, varying object types, and intricate environments.

ric Feature Extraction module (i.e., Ours CM) and without MF (i.e., Ours CM+GFE) showed a noticeable reduction in alignment accuracy compared to the complete model (Ours CM+GFE+MF). Specifically, the inclusion of the Geometric Feature Extraction module (Ours CM+GFE) improved alignment accuracy from 20.2% to 25.6%, and adding MF (Ours CM+GFE+MF) further increased accuracy by 0.2%. These results reveal the effectiveness of GFE in capturing both local and global geometric features. The ability of GFE to adapt to the geometric diversity of objects significantly enhances the alignment performance.

Effectiveness of Attention-Based Translation Estimation. Incorporating attention-based translation estimation has proven to be a crucial enhancement in our framework. An ablation study, where this feature was incrementally added, showed a notable improvement in model performance. Beginning with the basic model (Ours CM+GFE+MF), adding the attention-based translation estimation component (Ours CM+GFE+MF+ATE) resulted in a 1.2% increase in alignment accuracy, highlighting its significance. This attention-based mechanism leverages contextual and spatial relationships between objects, enabling more accurate position estimation. This approach not only improves the precision of translation predictions, but also contributes significantly to the overall robustness and applicability of the model across diverse scenarios.

Table 3. CAD model retrieval speed (in seconds) between traditional search method and TQGS. The traditional method processes each query by directly comparing vectors, while TQGS utilizes a more structured retrieval approach.

		Method	
Model Number	Feature Dimension	Traditional Search	TQGS
3000	256	0.7	0.1
	512	1.2	0.2
	1024	2.1	0.3
10000	256	1.2	0.1
	512	3.5	0.3
	1024	6	0.5

Effectiveness of Tri-Indexed Quantized Graph Search. To assess the efficiency of our Tri-Indexed Quantized Graph Search (TQGS) method, we conducted an ablation study comparing it to a traditional search method. The traditional method directly compares feature vectors, whereas TQGS employs a more structured retrieval approach. As shown in Table 3, with a feature dimension of 512 and a database of 3,000 CAD models, the traditional method takes 1.2 seconds to retrieve a relevant model, while TQGS reduces this time to 0.2 seconds, representing a 6x speedup. When

the number of models increases to 10,000, the traditional method requires 3.5 seconds, while TQGS completes the retrieval in just 0.3 seconds, resulting in an 11.7x speedup. This substantial reduction in retrieval time demonstrates the scalability and efficiency of TQGS. These results emphasize the critical role of the TQGS framework in significantly accelerating CAD model retrieval, especially in large-scale scenarios. The ablation study shows that TQGS not only reduces retrieval time but also maintains accuracy, making it a vital component for efficient and scalable CAD model retrieval.

Limitations. Despite its effectiveness, our method has some limitations. First, it assumes that the CAD models in the database are comprehensive and cover all possible objects in the images. If an object in the image lacks a corresponding CAD model, our method may not accurately represent it. Second, the method heavily depends on the quality of the 2D object recognition process. Therefore, any inaccuracies in object detection will directly impact overall performance. Future work will focus on addressing these limitations by expanding the CAD model database to cover a broader range of objects and improving the accuracy of the 2D object recognition process. Additionally, we aim to develop methods that can handle objects without corresponding CAD models, potentially through generative modeling or learning-based approximations.

5. Conclusion

In this work, we introduced CosCAD, a novel framework that leverages cross-modal CAD model retrieval and an attention-based method to significantly improve the accuracy of CAD model alignment with a single image. Our method outperforms existing techniques, demonstrating the effectiveness of integrating cross-modal information and estimating object locations by exploiting spatial correlations among objects. Various ablative studies further validate the importance of these components, emphasizing their contributions to the overall performance of the framework.

In future research, beyond addressing the above limitations and their improvements, another promising direction is joint retrieval and pose estimation, rather than the current method of processing each object individually. We believe that this strategy could improve the retrieval results and the accuracy of the pose estimation.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (No. T2322012, No. 62172218), and the Guangdong Basic and Applied Basic Research Foundation (No. 2022A1515010170).

References

[1] A. Avetisyan, M. Dahnert, A. Dai, M. Savva, A. X. Chang, and M. Nießner. Scan2cad: Learning CAD model alignment in RGB-D scans. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2614–2623. Computer Vision Foundation / IEEE, 2019. 8

[2] A. Avetisyan, A. Dai, and M. Nießner. End-to-end CAD model retrieval and 9dof alignment in 3d scans. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 2551–2560. IEEE, 2019. 8

[3] T. Binford. Survey of model-based image analysis systems. *The International Journal of Robotics Research,The International Journal of Robotics Research*, Mar 1982. 2

[4] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko. End-to-end object detection with transformers. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part I*, volume 12346 of *Lecture Notes in Computer Science*, pages 213–229. Springer, 2020. 2

[5] A. X. Chang, T. A. Funkhouser, L. J. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015. 1, 2

[6] H. Chen, M. Wei, Y. Sun, X. Xie, and J. Wang. Multi-patch collaborative point cloud denoising via low-rank recovery with graph constraint. *IEEE transactions on visualization and computer graphics*, 26(11):3255–3270, 2019. 2

[7] H. Chen, Z. Wei, X. Li, Y. Xu, M. Wei, and J. Wang. Repcd-net: Feature-aware recurrent point cloud denoising network. *International Journal of Computer Vision*, 130(3):615–629, 2022. 2

[8] H. Chen, Z. Wei, Y. Xu, M. Wei, and J. Wang. Imlovenet: Misaligned image-supported registration network for low-overlap point cloud pairs. In *ACM SIGGRAPH 2022 conference proceedings*, pages 1–9, 2022. 1

[9] Y. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu. UNITER: universal image-text representation learning. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 104–120. Springer, 2020. 2

[10] R. T. Chin and C. R. Dyer. Model-based recognition in robot vision. *ACM Comput. Surv.*, 18(1):67–108, 1986. 2

[11] C. B. Choy, D. Xu, J. Gwak, K. Chen, and S. Savarese. 3d-r2n2: A unified approach for single and multi-view 3d object reconstruction. In B. Leibe, J. Matas, N. Sebe, and M. Welling, editors, *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII*, volume 9912 of *Lecture Notes in Computer Science*, pages 628–644. Springer, 2016. 1

[12] A. Dai, A. X. Chang, M. Savva, M. Halber, T. A. Funkhouser, and M. Nießner. Scannet: Richly-annotated

3d reconstructions of indoor scenes. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2432–2443. IEEE Computer Society, 2017. 7

[13] K. Duan, S. Bai, L. Xie, H. Qi, Q. Huang, and Q. Tian. Centernet: Keypoint triplets for object detection. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6568–6577. IEEE, 2019. 2

[14] F. Engelmann, K. Rematas, B. Leibe, and V. Ferrari. From points to multi-object 3d reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 4588–4597. Computer Vision Foundation / IEEE, 2021. 2

[15] H. Fan, H. Su, and L. J. Guibas. A point set generation network for 3d object reconstruction from a single image. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 2463–2471. IEEE Computer Society, 2017. 1

[16] H. Fu, R. Jia, L. Gao, M. Gong, B. Zhao, S. J. Maybank, and D. Tao. 3d-future: 3d furniture shape with texture. *Int. J. Comput. Vis.*, 129(12):3313–3337, 2021. 1

[17] A. Gionis, P. Indyk, R. Motwani, et al. Similarity search in high dimensions via hashing. In *Vldb*, volume 99, pages 518–529, 1999. 5

[18] G. Gkioxari, J. Malik, and J. Johnson. Mesh R-CNN. *CoRR*, abs/1906.02739, 2019. 1

[19] T. Groueix, M. Fisher, V. G. Kim, B. C. Russell, and M. Aubry. A papier-mâché approach to learning 3d surface generation. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 216–224. Computer Vision Foundation / IEEE Computer Society, 2018. 1

[20] C. Gümeli, A. Dai, and M. Nießner. ROCA: robust CAD model retrieval and alignment from a single image. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 4012–4021. IEEE, 2022. 1, 3, 7, 8

[21] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 2, 3

[22] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 770–778. IEEE Computer Society, 2016. 2, 3

[23] H. Jegou, M. Douze, and C. Schmid. Product quantization for nearest neighbor search. *IEEE transactions on pattern analysis and machine intelligence*, 33(1):117–128, 2010. 5

[24] G. Jocher, A. Chaurasia, and J. Qiu. Ultralytics YOLO, Jan. 2023. 2

[25] W. Kuo, A. Angelova, T. Lin, and A. Dai. Mask2cad: 3d shape prediction by learning to segment and retrieve. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part III*, volume 12348 of *Lecture Notes in Computer Science*, pages 260–277. Springer, 2020. 1, 2, 7

[26] W. Kuo, A. Angelova, J. Malik, and T. Lin. Shapemask: Learning to segment novel objects by refining shape priors. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 9206–9215. IEEE, 2019. 2

[27] F. Langer, G. Bae, I. Budvytis, and R. Cipolla. SPARC: sparse render-and-compare for CAD model alignment in a single RGB image. In *33rd British Machine Vision Conference 2022, BMVC 2022, London, UK, November 21-24, 2022*, page 72. BMVA Press, 2022. 1, 3, 8

[28] J. Li, R. R. Selvaraju, A. Gotmare, S. R. Joty, C. Xiong, and S. C. Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9694–9705, 2021. 2

[29] L. H. Li, M. Yatskar, D. Yin, C. Hsieh, and K. Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019. 2

[30] L. H. Li, P. Zhang, H. Zhang, J. Yang, C. Li, Y. Zhong, L. Wang, L. Yuan, L. Zhang, J. Hwang, K. Chang, and J. Gao. Grounded language-image pre-training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10955–10965. IEEE, 2022. 2

[31] X. Li, X. Yin, C. Li, P. Zhang, X. Hu, L. Zhang, L. Wang, H. Hu, L. Dong, F. Wei, Y. Choi, and J. Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXX*, volume 12375 of *Lecture Notes in Computer Science*, pages 121–137. Springer, 2020. 2

[32] T. Lin, P. Dollár, R. B. Girshick, K. He, B. Hariharan, and S. J. Belongie. Feature pyramid networks for object detection. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 936–944. IEEE Computer Society, 2017. 1, 2, 3

[33] T. Lin, P. Goyal, R. B. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2999–3007. IEEE Computer Society, 2017. 2

[34] F. Liu and X. Liu. Voxel-based 3d detection and reconstruction of multiple objects from a single image. In M. Ranzato, A. Beygelzimer, Y. N. Dauphin, P. Liang, and J. W. Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 2413–2426, 2021. 7

[35] J. Lu, D. Batra, D. Parikh, and S. Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. M. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. B. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*

32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, pages 13–23, 2019. 2

[36] Y. A. Malkov and D. A. Yashunin. Efficient and robust approximate nearest neighbor search using hierarchical navigable small world graphs. *IEEE transactions on pattern analysis and machine intelligence*, 42(4):824–836, 2018. 5

[37] K. Maninis, S. Popov, M. Nießner, and V. Ferrari. Vid2cad: CAD model alignment using multi-view constraints from videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(1):1320–1327, 2023. 1, 3, 8

[38] N. Mu, A. Kirillov, D. A. Wagner, and S. Xie. SLIP: self-supervision meets language-image pre-training. In S. Avidan, G. J. Brostow, M. Cissé, G. M. Farinella, and T. Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXVI*, volume 13686 of *Lecture Notes in Computer Science*, pages 529–544. Springer, 2022. 2, 4

[39] Y. Nie, X. Han, S. Guo, Y. Zheng, J. Chang, and J. Zhang. Total3dunderstanding: Joint layout, object pose and mesh reconstruction for indoor scenes from a single image. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 52–61. Computer Vision Foundation / IEEE, 2020. 6, 7

[40] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. 2

[41] L. G. Roberts. *Machine Perception of Three-Dimensional Solids*. Outstanding Dissertations in the Computer Sciences. Garland Publishing, New York, 1963. 2

[42] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. S. Bernstein, A. C. Berg, and L. Fei-Fei. Imagenet large scale visual recognition challenge. *Int. J. Comput. Vis.*, 115(3):211–252, 2015. 1, 2

[43] H. Tan and M. Bansal. LXMERT: learning cross-modality encoder representations from transformers. In K. Inui, J. Jiang, V. Ng, and X. Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics, 2019. 2

[44] M. Tan, R. Pang, and Q. V. Le. Efficientdet: Scalable and efficient object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 10778–10787. Computer Vision Foundation / IEEE, 2020. 2

[45] M. Tatarchenko, A. Dosovitskiy, and T. Brox. Octree generating networks: Efficient convolutional architectures for high-resolution 3d outputs. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2107–2115. IEEE Computer Society, 2017. 1

[46] N. Wang, Y. Zhang, Z. Li, Y. Fu, W. Liu, and Y. Jiang. Pixel2mesh: Generating 3d mesh models from single RGB images. In V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, editors, *Computer Vision - ECCV 2018 - 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI*, volume 11215 of *Lecture Notes in Computer Science*, pages 55–71. Springer, 2018. 1

[47] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao. 3d shapenets: A deep representation for volumetric shapes. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, June 7-12, 2015*, pages 1912–1920. IEEE Computer Society, 2015. 1

[48] C. Zhang, Z. Cui, Y. Zhang, B. Zeng, M. Pollefeys, and S. Liu. Holistic 3d scene understanding from a single image with implicit representation. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 8833–8842. Computer Vision Foundation / IEEE, 2021. 2

[49] R. Zhang, Z. Guo, W. Zhang, K. Li, X. Miao, B. Cui, Y. Qiao, P. Gao, and H. Li. Pointclip: Point cloud understanding by CLIP. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 8542–8552. IEEE, 2022. 2

[50] L. Zheng, C. Wang, Y. Sun, E. Dasgupta, H. Chen, A. Leonardis, W. Zhang, and H. J. Chang. Hs-pose: Hybrid scope feature extraction for category-level object pose estimation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 17163–17173. IEEE, 2023. 6

[51] H. Zhou, H. Chen, Y. Feng, Q. Wang, J. Qin, H. Xie, F. L. Wang, M. Wei, and J. Wang. Geometry and learning co-supported normal estimation for unstructured point cloud. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13238–13247, 2020. 1