International Journal of Geo-Information

MDPI

Article

View-Invariant 3D Building Retrieval with Topological Perception-Guided Feature Fusion

Xinwen Zhang [1], Yuan Ding [2,*], Yi Lu [2], Xiaoping Rui [1], Hua Shao [3] and Jin Zhu [4]

[1] School of Earth Sciences and Engineering, Hohai University, Nanjing 211000, China; 231309020037@hhu.edu.cn (X.Z.); ruixp@hhu.edu.cn (X.R.)
[2] College of Geography and Remote Sensing, Hohai University, Nanjing 211100, China; yi.lu@hhu.edu.cn
[3] School of Geomatics Science and Technology, Nanjing Tech University, Nanjing 211800, China; shaohua@njtech.edu.cn
[4] School of Geography Science and Geomatics Engineering, Suzhou University of Science and Technology, Suzhou 215009, China; zhujin@usts.edu.cn
* Correspondence: dingyuanhhu@hhu.edu.cn

Abstract

The increasing availability of 3D building models in digital-city applications has made scalable and accurate 3D building model retrieval essential. However, existing methods often struggle to capture the global structure of building models and to achieve stable retrieval results under viewpoint variations. To address these challenges, we propose a topological-perception-guided feature fusion framework with two complementary fusion schemes for different computational budgets. The topological perception features capture global structure and provide relatively viewpoint-stable information, and they are respectively fused with traditional features and deep features for low- and high-compute-budget scenarios. In addition, the topological perception features guide view selection and view grouping to improve retrieval stability. Experiments show that the traditional-feature fusion scheme improves retrieval accuracy by 8.0–25.7 percentage points, while the deep fusion scheme outperforms Multi-view Convolutional Neural Networks (MVCNNs) and Group-View Convolutional Neural Networks (GVCNNs) by 1.2 and 4.0 percentage points, respectively. These results suggest that incorporating topological perception as guidance for feature fusion strengthens global structural representation and supports viewpoint-invariant retrieval for architecturally complex building models.

Keywords: 3D building model retrieval; topological perception; viewpoint invariance; feature fusion

Check for updates

Academic Editor: Wolfgang Kainz

Received: 8 January 2026
Revised: 2 March 2026
Accepted: 24 March 2026
Published: 30 March 2026

Copyright: © 2026 by the authors. Published by MDPI on behalf of the International Society for Photogrammetry and Remote Sensing. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

1. Introduction

Driven by the continuous advancement of digitalization and 3D modeling technologies, 3D building models have been widely applied in various fields, including urban planning [1], architectural design [2], virtual reality [3], and cultural heritage preservation [4]. In this context, 3D building model retrieval, a technology that automatically finds models similar to a query model from large-scale model repositories, has become an indispensable tool in fields such as digital cities and smart buildings. Efficient and accurate 3D model retrieval not only enhances the efficiency of architectural design and planning but also significantly supports construction projects by ensuring better decision-making [5–7]. However, with the increasing scale and complexity of building models, the challenge of quickly and accurately retrieving the target model from large model databases has become significant [8].

A key factor influencing the performance of 3D building model retrieval lies in how effectively model features can be represented and compared. In particular, building models are characterized by strong structural coherence (e.g., massing composition, symmetry, and hierarchical part–whole structure), and retrieval aims to recognize such buildings by comparing these global structural features across models. Therefore, effective extraction of global features is crucial for building model retrieval. Existing retrieval methods for 3D models primarily rely on 3D geometric features and 2D view features [9–13]. Traditional approaches, in particular, often employ single handcrafted descriptors [10,11]. In practice, commonly used 3D descriptors often focus on extracting local geometric details [14], which challenges the stable representation of the global features of building models, thereby reducing the accuracy of retrieval results.

Traditional 3D model retrieval methods based on traditional features can improve accuracy by combining complementary descriptors, thereby capturing both global geometry and local details of building models [15]. Deep feature-based retrieval methods, in contrast, rely on neural networks to automatically learn representations for matching, reducing dependence on manually designed descriptors and providing a certain level of global features [16]. However, both approaches are predominantly model-driven rather than perception-driven: they do not explicitly incorporate perception-inspired global and viewpoint-stable characteristics, and their performance can be sensitive to the randomness of view selection. As a result, retrieval outcomes may vary noticeably when the viewpoint changes, especially for buildings with complex structures (e.g., concavities or holes), where small viewpoint perturbations can cause substantial appearance differences in projected views [17]. In this paper, we use viewpoint invariance to describe the stability of retrieval results under viewpoint variations, i.e., the retrieval results should not fluctuate markedly with different selections of side views. Therefore, how to describe the features of 3D building models from the perspective of global and viewpoint-stable characteristics in visual perception has become a key issue in improving the accuracy and robustness of 3D building model retrieval.

Topological perception, as a form of visual perception characterized by global priority and invariance, can address the shortcomings of existing methods in expressing global visual perceptual features [18]. Although a prior study has attempted to apply topological perception features to 3D model retrieval, it mainly uses a single topological perception feature as a standalone representation, which remains insufficient for complex models [19]. In response to this limitation, we develop a topological-perception-guided feature fusion framework that leverages the Quotient Distance–Angle Histogram (QDAH) to inject global and relatively viewpoint-invariant structural cues into 3D building model retrieval. The framework comprises two complementary fusion schemes tailored to low- and high-compute-budget settings, respectively. In low-compute-budget settings, QDAH is fused with conventional handcrafted descriptors and also guides representative-view selection to enhance viewpoint robustness. In high-compute-budget settings, QDAH is fused with deep multi-view representations and further supports view grouping prior to multi-view aggregation, thereby improving retrieval stability for architecturally complex building models.

Moreover, the topological perception features captured by QDAH (e.g., holes and concavities) are integral to specific 3D building modeling and reconstruction pipelines, including template constraint parameterization and rule-based template construction. Nevertheless, they are not guaranteed to uniquely identify individual buildings and do not provide sufficient metric information for detailed reconstruction; thus, they should be regarded as auxiliary high-level perceptual priors, structural constraints, or consistency checks, rather than standalone reconstruction descriptors.

The remainder of this paper is organized as follows: Section 2 reviews mainstream 3D building model retrieval methods and their limitations, and discusses topological perception in retrieval, including its distinction from topological relations and related studies. Section 3 introduces the computation of topological perception features and matching strategies, and provides a detailed explanation of the proposed topological-perception-guided feature fusion framework and its two fusion schemes. Section 4 evaluates the effectiveness of the proposed method through a comparative analysis with other related algorithms. Section 5 discusses the experimental findings, and Section 6 concludes the study and outlines directions for future research.

2. Related Works

2.1. Retrieving 3D Models Using Traditional Features

Traditional feature-based methods play a crucial role in 3D model retrieval, but most focus on the extraction of local features, which provide limited support for capturing the global features that are critical for building models. Depending on the feature extraction methods, traditional approaches can be broadly classified into categories such as geometry-based statistical features, view-based features, and hybrid feature-based methods.

Geometry-based statistical feature methods describe the geometric distribution of 3D shapes by constructing statistical histograms, focusing primarily on the extraction of local geometric features. For example, the Radius–Normal Histogram (RNH) method constructs a rotation-invariant statistical descriptor by calculating the radial and normal angle between the surface sampling points of the model, effectively improving the robustness of retrieval [10]. Kazhdan et al. [11] used the Spherical Harmonics (SPH) descriptor to represent 3D models, making the descriptor invariant to rotation and effectively addressing the issue of rotational interference during model matching. Nevertheless, although these methods handle rotation well, they still rely mainly on local measurements, which makes it difficult to capture the global geometric relationships of complex building models and may limit retrieval performance for highly similar shapes.

View-based feature methods project 3D models onto a 2D plane and extract features from the projection images for retrieval. For example, simplified projection methods directly use silhouettes of the 3D model as a feature, enabling fast model matching [12]. The Light Field Descriptor (LFD) generates multi-view projections by enclosing the model with a dodecahedron to capture richer view information [13]. However, because these features are extracted from specific viewpoints, retrieval performance can vary with view selection. For building models, limited views may fail to describe the global structure accurately, and random viewpoint selection can further reduce retrieval stability and accuracy.

To describe both geometric structure and detailed information more comprehensively, some studies have proposed hybrid feature-based retrieval methods. For example, Bag-of-Features SIFT (BoF-SIFT) captures local features using the SIFT algorithm and fuses them through a Bag-of-Features (BoF) model to enhance retrieval accuracy for complex models [20]. Tu [21] proposed a multi-feature fusion method that fuses HOG, Zernike-moment, and Fourier features for projection images, and further uses an online projection learning algorithm to fuse multi-view features, improving retrieval performance and feature discrimination. However, although these methods fuse multiple features and can improve retrieval performance, the fused representation can still be dominated by local or view-dependent information. This makes it difficult to capture the overall structure of building models, limiting discrimination among highly similar buildings.

In summary, traditional feature-based retrieval methods remain practically relevant in many retrieval pipelines due to their training-free nature and low computational overhead. However, they are limited in capturing stable global features of complex building models,

which may lead to degraded performance when distinguishing highly similar buildings. Accordingly, with the rapid development of deep learning in fields such as computer vision [22–24], increasing attention has turned to deep feature-based retrieval to achieve stronger global representation.

2.2. Retrieving 3D Models Using Deep Features

Building on this trend, deep learning has become an important direction for 3D model retrieval. Compared with traditional feature-based methods, deep networks learn multi-level feature representations from data, which can capture both the overall structure and fine details of 3D models and improve retrieval performance for 3D building models. Deep feature-based 3D model retrieval methods are commonly grouped into view-based, voxel-based, and point cloud-based methods. This paper focuses on view-based 3D building model retrieval, where a 3D model is represented by multiple projected 2D views and similarity is computed by learning view features and fusing them across views.

In this view-based setting, deep learning has advanced view-based 3D model retrieval by improving view feature learning and similarity measurement. Convolutional Neural Networks (CNNs) learn high-level semantic features from images, which helps describe building appearance and improves view-based retrieval performance [25]. Representative attention-based models, such as Adaptive Attentional Landmark Image Retrieval (AALIR) and the Convolutional Neural Network Attention Retrieval framework (CNNAR), further enhance discriminative view representations for projected views [26,27]. These studies demonstrate the effectiveness of learned features on 2D views and provide useful motivation for view-based 3D retrieval. However, directly relying on view features is still insufficient for 3D building model retrieval, because projected views can vary noticeably under viewpoint variations, and retrieval performance can be sensitive to which views are selected.

Multi-view fusion is a key step in view-based 3D model retrieval, since view features from different viewpoints must be fused into a single representation for matching. Multi-view Convolutional Neural Networks (MVCNNs) extract view features using a convolutional network and fuse them across views to obtain a global representation of the overall structure, which has become a foundational framework for 3D model classification and retrieval [28]. To improve the utilization of multi-view information, Group-View Convolutional Neural Networks (GVCNNs) introduce view grouping to strengthen information exchange among related views and improve representation for complex models [29]. To further optimize the fusion of multi-view features, Dynamic Routing Convolutional Neural Networks (DRCNNs) introduce a dynamic routing layer, effectively improving the performance of feature fusion [30]. However, despite these advances, retrieval performance can still be sensitive to view selection and viewpoint variations. This issue becomes more pronounced for models with complex geometric structures, where small viewpoint changes may lead to noticeable differences in projected views and cause fluctuations in retrieval accuracy.

Overall, the above limitations suggest that further improving retrieval stability requires a guiding feature that remains stable under viewpoint variations. From the perspective of human vision, topological perception features provide global and relatively viewpoint-stable information. Motivated by this, we next review topological perception theory and its relevance to 3D model retrieval.

2.3. Topological Perception and Topological Relations in 3D Retrieval

Topological perception theory describes how humans perceive objects by prioritizing topological perception features [31]. It suggests that features such as holes, concavities,

and convexities convey global and relatively invariant information, and can be processed earlier than local geometric details in visual perception [32]. Topological perception theory has motivated perception-inspired studies in computer vision and visual design, and has been applied in areas including object recognition [33], image saliency detection [34], and product design [35].

Beyond these applications, similar perception-driven ideas also appear in GIScience and cartography in the context of landmark building depiction and visualization. For instance, Ti et al. proposed an automatic approach to generate outline-based representations of landmark buildings with distinctive shapes to improve recognizability in map depiction [36]. Elias and Paelke discussed user-centered design principles for landmark visualization, showing how abstraction and visual saliency support the perception and interpretation of landmark buildings in map-based applications [37]. Although these studies focus on cartographic representation rather than 3D model retrieval, they similarly emphasize perceptually salient global structure, providing a broader context and additional support for introducing topological perception features as globally informative and viewpoint-stable features in our retrieval framework.

It is important to distinguish retrieval based on topological perception from retrieval based on topological relations. Topological relations describe connectivity properties among geometric elements or components in 3D models, such as adjacency, connectivity, and closure [38–40]. In contrast, topological perception is a theory of visual cognition and focuses on topological perception features such as holes and concavities, which convey global and relatively invariant information [32,41]. Although both involve connectivity-related notions, we use topological perception features as a perception-inspired representation for retrieval, rather than explicitly modeling relational topology between model components.

While a prior study has introduced topological perception features for 3D model retrieval, it primarily relies on a single type of such features as a standalone representation [19], which limits its expressiveness for complex structures. To address this, we propose a unified topological-perception-guided fusion framework with two fusion schemes to accommodate different levels of computational resources: one scheme fuses topological perception features with traditional features, while the other scheme fuses them with deep features. Together, the two schemes enhance global structural representation and improve retrieval stability under viewpoint variations, especially for architecturally complex building models.

3. Method

3.1. Overview of the Method

From a methodological perspective, 3D model retrieval methods can be broadly grouped into two paradigms with markedly different computational budgets. Handcrafted-descriptor-based retrieval is typically lightweight and training-free, relying on predefined geometric or view statistics and explicit similarity measures. By contrast, deep-learning-based retrieval often yields stronger discriminative power but is more compute-intensive due to offline model training and multi-view feature extraction at query time [42]. Motivated by this budget gap, we design two complementary fusion schemes to accommodate different retrieval settings. In our implementation, these two schemes correspond to Fusion Algorithm 1 (traditional-feature fusion) and Fusion Algorithm 2 (deep-feature fusion), respectively.

As shown in Figure 1, the method begins by generating a set of 2D projections using a nine-directional projection method, from which topological perception features are extracted using QDAH. These topological perception features provide global and relatively viewpoint-stable structural information, serving as guidance for subsequent feature fusion and view management.

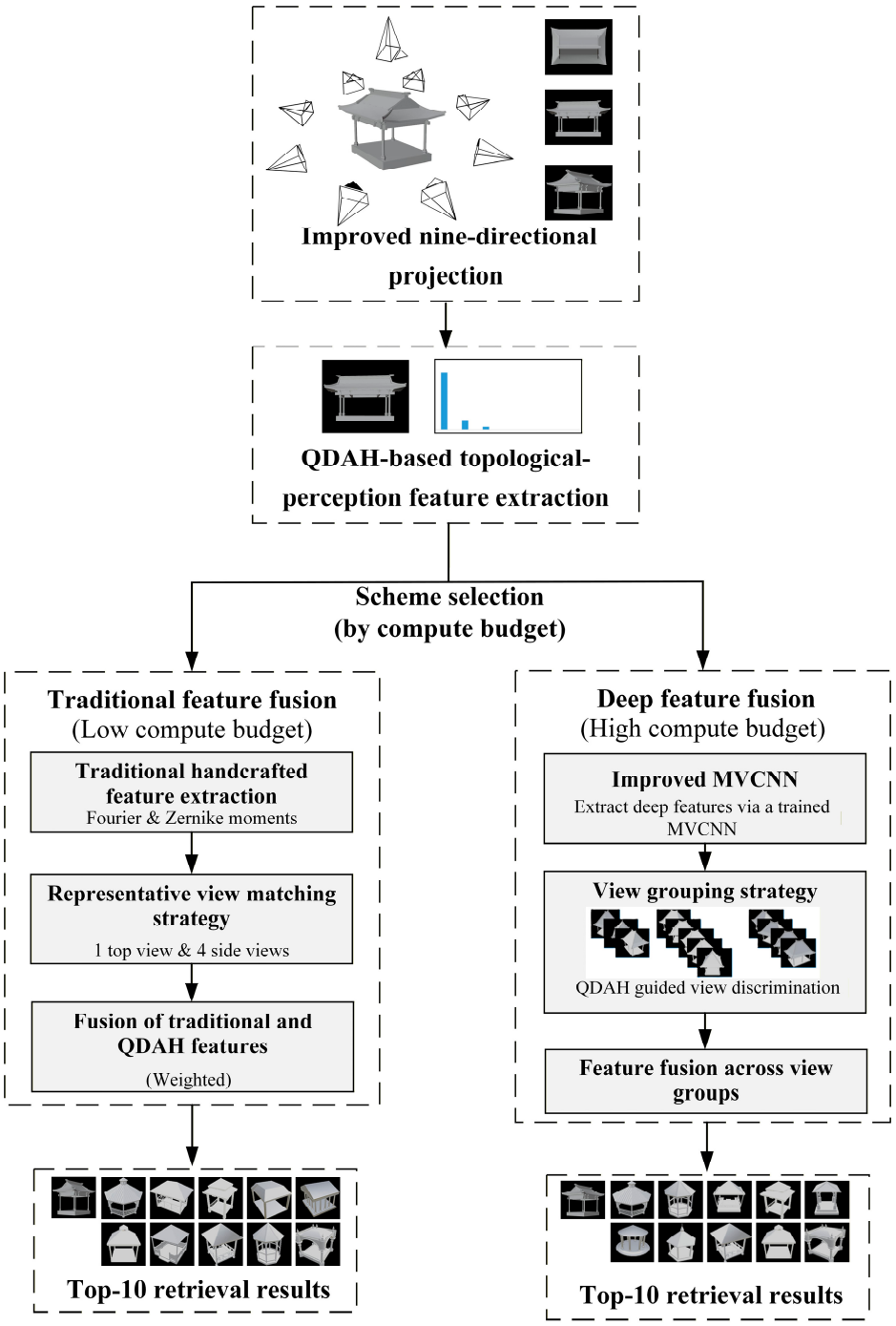


Figure 1. Overall framework of the proposed 3D building model retrieval method based on topological-perception-guided feature fusion.

Depending on the available computational resources, the framework then follows Fusion Algorithm 1 or Fusion Algorithm 2 for retrieval. The traditional-feature fusion scheme is designed for budget-constrained settings. It stabilizes handcrafted matching by using QDAH to guide representative-view selection, and then combines QDAH with conventional handcrafted descriptors through weighted fusion. The deep-feature fusion scheme targets higher compute budgets. It integrates QDAH-guided view grouping into an improved MVCNN, so that multi-view features are aggregated across grouped views via an attention-based mechanism.

Each scheme produces a retrieval ranking for the query model by combining QDAH with either traditional descriptors or learned deep representations. Detailed descriptions of

the projection method, feature extraction techniques, and fusion schemes are provided in Sections 3.2–3.4.

3.2. Extracting Topological Perception Features from Projected Views

To extract topological perception features from 3D building models, we first generate a set of consistent 2D projected views using the nine-directional projection setting. Specifically, one virtual camera is placed above the model to capture the top view, and eight cameras are placed around the model at 45° intervals to capture side views. All cameras are directed toward the centroid of the model, with the side cameras distributed on a plane that passes through the model's centroid and is parallel to the ground plane (see Figure 2). Through this projection system, each building model generates one top view and eight side views.

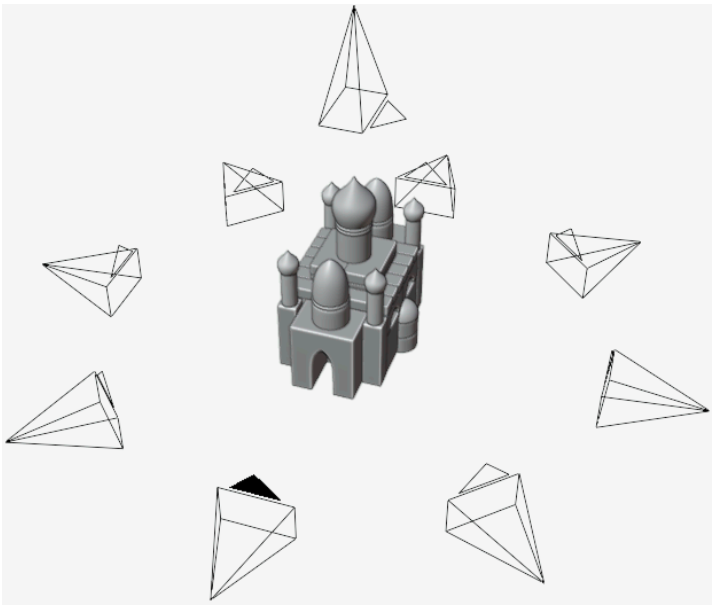


Figure 2. Camera configuration in the nine-directional projection method for multi-view generation.

Based on these projected views, we compute QDAH to capture perceptually salient topological features, such as holes and concavities, in a compact histogram descriptor. Here, "perceptually salient" is used in a method-level sense and refers to structures that yield stable and discriminative QDAH distributions across projected views, rather than being interpreted as a subjective notion of human perception or implemented as a manual equal-weighting of individual holes or concavities. We do not impose additional explicit conditions, such as size thresholds or spatial-dominance weighting. In the QDAH-based representation, saliency is implicitly encoded by aggregating quotient-distance statistics over multiple views, which emphasizes view-consistent and discriminative distributions. It is explicitly strengthened by the angle refinement, which enhances sensitivity to closure-related differences and reduces ambiguity between visually similar structures, such as closed holes and concave non-closed shapes.

Following the established procedure in prior work [19,43], QDAH is constructed by first computing a quotient-distance-based histogram (QDH) and then integrating angle distribution information to better distinguish visually similar structures. The quotient distance between two pixels x and y is defined as

$$qd(x, y) = d^*(x, y)/d_1(x, y) \quad (1)$$

where $d_1(x, y)$ denotes the Euclidean distance and $d^*(x, y)$ denotes the corresponding distance measured in the topological space [43]. The QDH is then obtained by accumulating $qd(x, y)$ into histogram bins, as given in (2):

$$f(i) = \sum_{x=1}^{n} \sum_{y=x+1}^{n} I(qd(x, y) \in B(i)) \tag{2}$$

where $B(i)$ is the range of the *i-th* bin, and $I(\cdot)$ is the indicator function. The subsequent angle-based refinement follows [19] to reduce ambiguity between visually similar structures (e.g., holes versus concave non-closed shapes). Figure 3 provides a schematic illustration showing that the aggregation of QDH across views, together with the subsequent angle refinement, encodes topology-related differences into QDAH distributions, thereby improving the separability between representative structures (e.g., closed holes versus concave non-closed shapes).

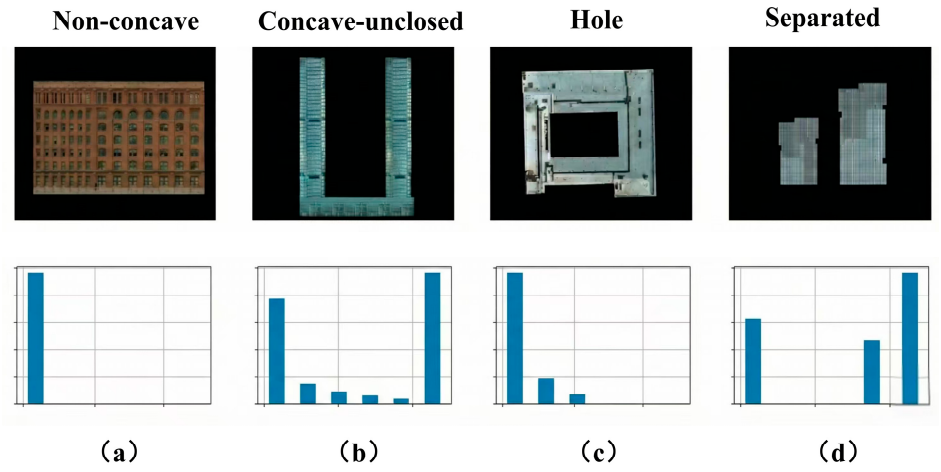


Figure 3. Average quotient distance–angle histogram (QDAH) examples for representative topological structures, showing projected views and corresponding mean QDAH distributions for (**a**) non-concave, (**b**) concave-unclosed, (**c**) hole, and (**d**) separated cases.

Finally, to accommodate different computational budgets, the projected views and QDAH are used differently across the two fusion schemes: the traditional-feature fusion scheme uses representative-view selection and weighted combination, whereas the deep-feature fusion scheme uses QDAH-guided view grouping before multi-view aggregation. Details are provided in Sections 3.3 and 3.4.

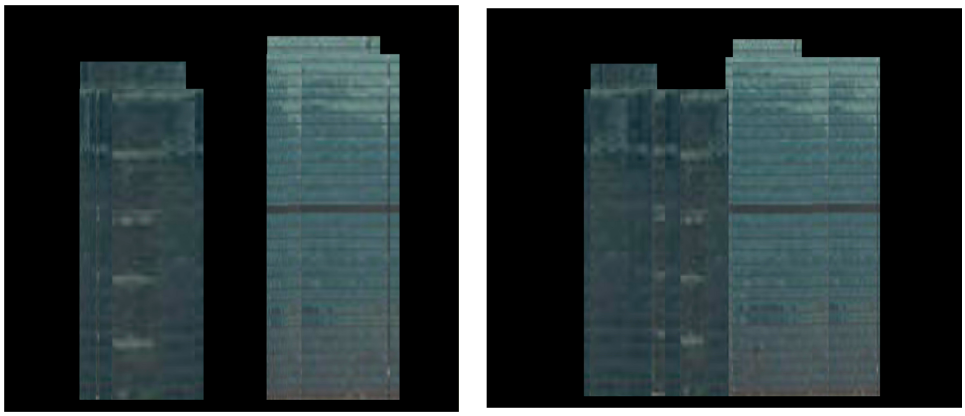
3.3. Fusing Traditional Features with Topological Perception

3.3.1. Matching Strategy for Traditional Feature-Based Algorithms

Traditional feature-based retrieval typically measures similarity by matching projected views between two models. A common solution is to compute a complete view-to-view distance matrix and then solve an optimal assignment problem using the Hungarian algorithm. While effective in general multi-view matching, this full matching strategy is not always desirable for building retrieval, where many side views are highly redundant and small viewpoint perturbations can yield multiple near-equivalent assignments. As a result, the matching outcome may become sensitive to view sampling, and the computational cost increases with the number of views. For completeness, we report a comparison between Hungarian matching and representative-view matching in Section 4.2; here, we focus on the rationale and procedure of the representative-view strategy adopted in our traditional-feature fusion scheme.

To obtain a lightweight yet stable matching procedure for the traditional-feature fusion scheme, we follow the representative-view selection strategy introduced in prior work [19]. The key idea is to use QDAH features to select a compact set of informative views before matching. Considering that adjacent side views often exhibit strong visual similarity for building models [44], we retain one top view and select four side views at 90° intervals for traditional feature extraction and similarity computation, which balances descriptiveness and efficiency.

This motivates distinguishing concave and non-concave cases when selecting the representative side view: for models with concave or disconnected structures, concavity may only appear in some viewing directions, whereas truly non-concave models remain non-concave under all projected side views (see Figure 4).



(**a**) (**b**)

Figure 4. View-dependent variation in topological perception features in projected side views for a disconnected structure. (**a**) A side view where the separation between components is clearly visible; (**b**) another side view of the same model where the separation is occluded in the projection, making the structure appear as a single non-concave shape.

We then select the representative side view from the four candidates using QDAH-based discrimination (see Figure 5). Specifically, QDAH features are extracted from each side view and compared with a set of predefined topological categories; the category with the smallest topological distance is used to determine the closest structure class for that view [45]. Therefore, if concave side views exist, the representative side view is selected among the concave candidates; otherwise, it is selected among the non-concave candidates by choosing the one with the smallest topological distance to its closest predefined category.

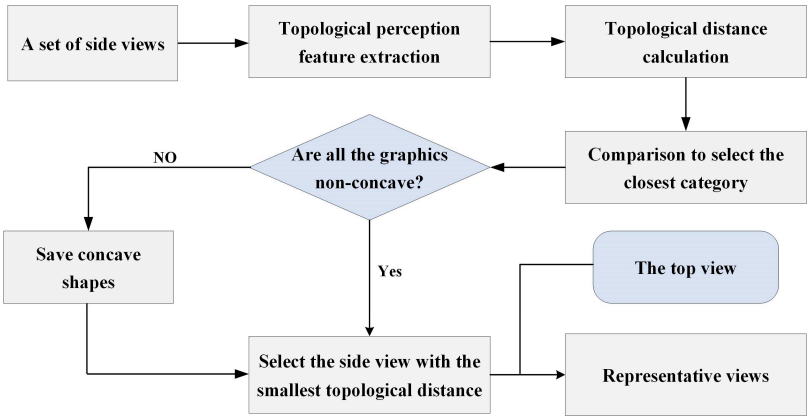


Figure 5. Workflow for representative view selection based on topological perception features. The optimal side view is selected via concavity classification and minimum topological distance.

After selecting the representative side view, we compute view-level distances using both handcrafted descriptors and QDAH features. Specifically, Fourier and Zernike descriptors are extracted from the top view and the representative side view, and their distances are fused as the first-layer distance (Section 3.3.2). In parallel, the QDAH distance is computed between the corresponding views. The final retrieval distance is obtained by a second-layer fusion of the handcrafted distance and the QDAH distance, and models are ranked accordingly.

3.3.2. Describing and Fusing Traditional Feature Descriptors

Fourier descriptors [46] are often represented using low-frequency coefficients to capture global shape characteristics, whereas Zernike moments [47] typically use higher-order moments to encode richer shape details. Following prior empirical settings [48], we extract features from Fourier descriptors and Zernike descriptors by retaining the first 10 low-frequency Fourier coefficients and 35 higher-order Zernike moments. For the 2D views of 3D building models, this study employs the L1 norm of vectors to measure the Fourier descriptors and Zernike descriptors, as shown in Equations (3) and (4):

$$D_F = \sum_{a=1}^{10} |F_1(a) - F_2(a)| \tag{3}$$

$$D_z = \sum_{b=1}^{35} |Z_1(b) - Z_2(b)| \tag{4}$$

where $F(a)$ and $Z(b)$ correspond to the Fourier transform features and Zernike moment features of the 2D views, respectively. D_F and D_Z represent the similarity results between the views of the query model and the views in the model library, obtained using the Fourier descriptors and Zernike descriptors, respectively. These two descriptors are fused to construct the first-layer fused descriptor D_1 in this algorithm, as shown in Equation (5):

$$D_1 = \alpha \cdot \frac{D_F - \min(D_F)}{\max(D_F) - \min(D_F)} + \beta \cdot \frac{D_Z - \min(D_Z)}{\max(D_Z) - \min(D_Z)} \tag{5}$$

where α and β are weighting coefficients that balance the contributions of the two descriptors. Prior to fusion, each descriptor is normalized to mitigate the effect of scale differences. Here, min and max denote the minimum and maximum values within each descriptor vector, respectively, used for min-max normalization. This normalization ensures that neither descriptor dominates the fused result due to scale discrepancies. After min–max normalization, we set $\alpha = \beta = 0.5$ to assign equal contributions to the Fourier and Zernike descriptors, thereby avoiding potential bias caused by arbitrary weighting.

To overcome the limitations of a single descriptor in 2D view feature extraction, this study combines the first-layer fused descriptor with the QDAH descriptor and performs a secondary fusion by setting appropriate weights. This approach constructs a more comprehensive and refined 2D view feature descriptor. The second-layer fused descriptor D is defined in Equation (6):

$$D = \omega D_1 + (1 - \omega) D_Q, \omega \in (0, 1] \tag{6}$$

where D_Q represents the similarity calculation results of the QDAH descriptor between model views, and ω denotes the weight of the first-layer fused descriptor. A smaller D indicates higher similarity between the query and the target model. Here, ω controls the trade-off between the first-layer fused descriptor D_1 and the QDAH-based similarity term D_Q in the second-layer integration. The value of ω is determined empirically via the parameter study described in Section 3.3.3 and shown in Figure 6.

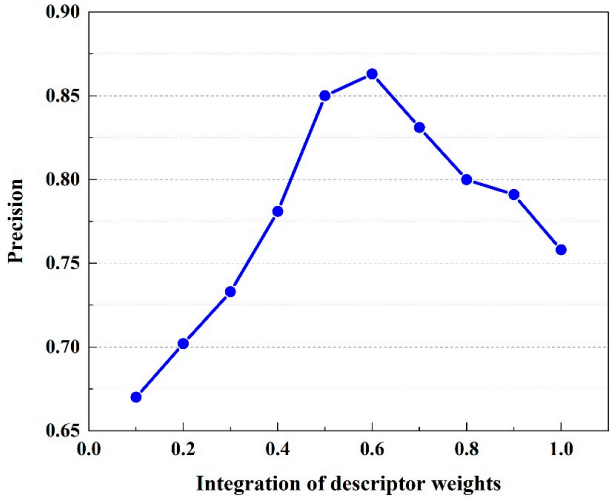


Figure 6. Effect of the fused descriptor weight on retrieval precision in traditional feature fusion.

3.3.3. Calculating Weight for Fused Descriptors

To validate the impact of ω in the fused descriptor on algorithm performance, this study conducted related experiments using a self-constructed 3D building model database (detailed in Section 4). Specifically, we varied ω from 0.1 to 1.0 with a step size of 0.1 and computed retrieval precision for each setting, as shown in Figure 6.

The retrieval precision increases with the fused descriptor weight from 0 to 0.6, but decreases gradually when the weight exceeds 0.6 up to 1.0. The best performance is obtained when the fused descriptor weight lies in 0.5–0.7, with the peak at 0.6. Accordingly, we use 0.6 as the fused descriptor weight in subsequent experiments.

3.4. Fusing Deep Features with Topological Perception

3.4.1. Improved MVCNN

MVCNN improves the recognition performance of 3D objects by fusing information from multiple views. To support the deep-feature fusion scheme, this study adopts an improved MVCNN architecture, as illustrated in Figure 7.

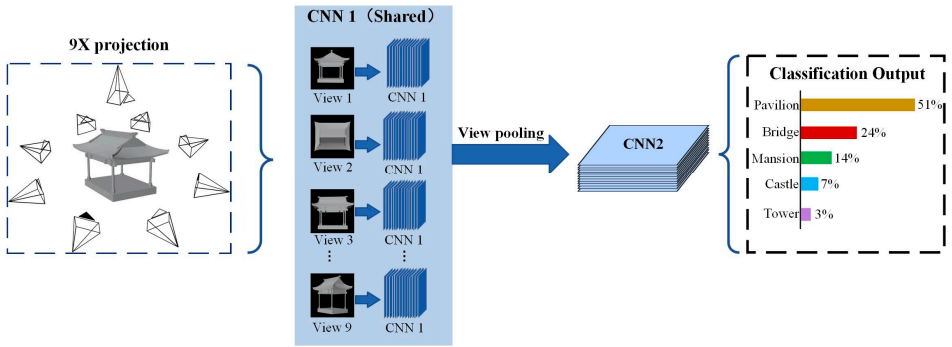


Figure 7. Architecture of the improved MVCNN with GoogLeNet-based multi-view feature extraction and pooling.

The improved MVCNN architecture comprises three key components: the first CNN layer, the view pooling layer, and the second CNN layer. Specifically, at the input stage, the 3D building model is first subjected to a nine-directional projection to generate a set of views. Then, each view image is input individually into the first convolutional neural network to extract features, with all branches of the network sharing the same weight parameters. To enhance the network's feature representation capability, this study replaces

the first 5 layers of VGG-M, traditionally used in MVCNN, with the first 21 layers of GoogLeNet. The features extracted from each view are fused through the view pooling layer and aggregated into a compact global representation of the 3D building model. Finally, the fused features are input into the second convolutional neural network for classification.

The improved MVCNN architecture is designed to strengthen feature representation and enable more effective multi-view aggregation, which serves as the deep feature extractor in the proposed deep-feature fusion scheme.

3.4.2. Matching Strategy for Deep Feature-Based Algorithms

Under the deep-feature fusion scheme, we leverage richer multi-view inputs to support representation learning under a higher compute budget. To this end, this paper proposes a group-view-based matching strategy, which optimizes the feature extraction and matching process by reasonably grouping multi-view data, thereby improving retrieval accuracy and performance. The view grouping module designed in this study is shown in Figure 8.

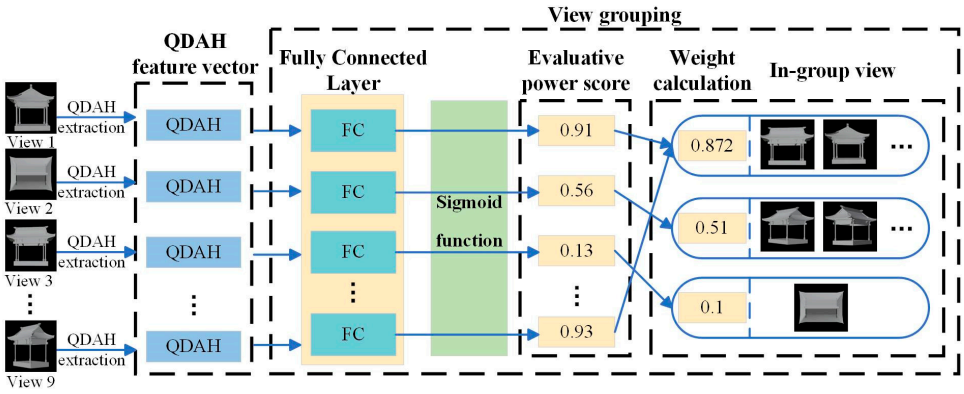


Figure 8. QDAH-guided view grouping for deep-feature fusion, where per-view QDAH descriptors are mapped to discriminative scores to partition views into N groups and perform weighted aggregation.

This module first generates views using the nine-directional projection method and extracts the corresponding QDAH feature vectors. Subsequently, the feature vectors are input into a fully connected layer for processing. Finally, each feature vector is mapped to a single value to evaluate the discriminative power of the view, and the calculation formula is as follows:

$$\xi(I_i) = sigmoid(log(abs(O_{I_i}))) \tag{7}$$

where O_{I_i} represents the output of the fully connected layer. To prevent the output from approaching 0 or 1, absolute value and logarithmic operations are added. The function *sigmoid* maps numbers to the range (0, 1), and I_i denotes a view in the view set $\{I_1, I_2, \ldots I_9\}$.

After obtaining the discriminative power scores, the views are divided into N equal intervals based on their scores, and the views within the same interval are grouped together. The features for each group are calculated using Equations (8) and (9).

$$D(G_j) = \frac{\sum_{i=1}^{N} \lambda_i D_{I_i}}{\sum_{i-1}^{N} \lambda_i} \tag{8}$$

where G_j represents the group, D_{I_i} denotes the deep feature extracted from the I_i view, and λ_i is the weighting factor, defined as follows:

$$\lambda_i = \begin{cases} 1 & I_i \in G_j \\ 0 & I_i \notin G_j \end{cases} \tag{9}$$

This study assumes that the view features within the same group are similar, and thus the group feature D_{G_j} for group G_j can be calculated by averaging. The feature information of different groups varies: groups with higher discriminative power are considered to contain more feature information and are therefore assigned higher weights, while groups with lower discriminative power are given lower weights to reduce their interference in model retrieval. The weight calculation for each group is shown in Equation (10):

$$\xi(G_j) = \frac{\text{Ceil}(\xi(I_k) \times |G_j|)}{|G_j|} \quad I_k \in G_j \tag{10}$$

where Ceil represents the ceiling function, and $|G_j|$ denotes the number of members in group G_j. The final model feature descriptor $D(S)$ is calculated using Equation (11):

$$D(S) = \frac{\sum_{j=1}^{M} \xi(G_j) D(G_j)}{\sum_{j=1}^{M} \xi(G_j)} \tag{11}$$

After determining the calculation method for the grouping module, another critical issue is defining the number of view groups for the model. The number of view groups is crucial for the performance of 3D model retrieval. In extreme cases, if the number of view groups is set to 1, all views are assigned to the same group, making the feature descriptor equivalent to 3D average pooling, which results in limited discriminative power and lower retrieval accuracy. Conversely, if the number of groups equals the number of views, the classification of views becomes overly detailed, and redundant information from similar views can lead to a decline in retrieval performance.

To identify the optimal number of view groups, this study conducted model retrieval experiments using retrieval accuracy as the evaluation criterion. The number of groups was sequentially set from 1 to 9, with the number of training epochs fixed at 10. The experimental results are shown in Figure 9.

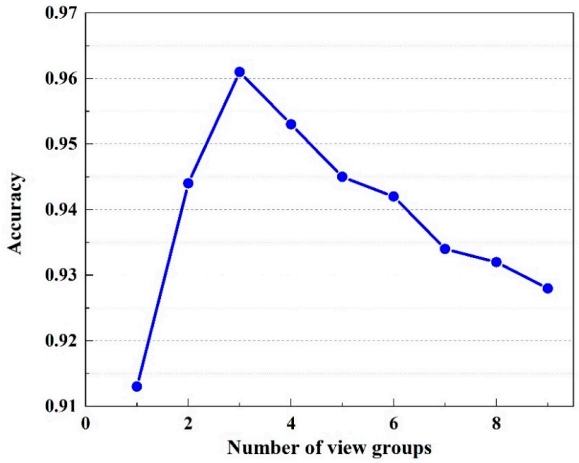


Figure 9. Retrieval accuracy of the deep feature fusion model under varying numbers of view groups.

Figure 9 demonstrates that the number of model view groups has a significant impact on retrieval accuracy, with a difference of up to 4.4 percentage points between the best and worst results. In the two extreme cases, the grouping loses its classification significance, resulting in the poorest retrieval performance. When increasing the number of groups from 1 to 3, retrieval accuracy rises rapidly, and then gradually decreases as the grouping becomes overly fine-grained. This may be related to the number of categories used in the classification of model views based on topological perception features. Ultimately, the

optimal result is achieved when the number of groups is set to 3, which will be used in subsequent experiments.

3.4.3. Fusing Deep Features with QDAH-Guided View Grouping and Attention Mechanism

The visual attention mechanism enables humans to optimize the allocation of attention when processing large amounts of information, focusing on important regions while ignoring irrelevant details. Inspired by this mechanism, this study incorporates a view grouping matching strategy into the deep feature fusion process and designs an attention allocation mechanism based on deep networks.

Specifically, a set of views is first generated through the nine-directional projection method and grouped into several subsets using the view grouping strategy. These grouped view features are then input into the improved MVCNN network, allowing the network to accurately focus on important view information while disregarding less important parts. This process optimizes the importance distribution of features both within and across groups, thereby enhancing the model's representation capability and recognition performance, as shown in Figure 10:

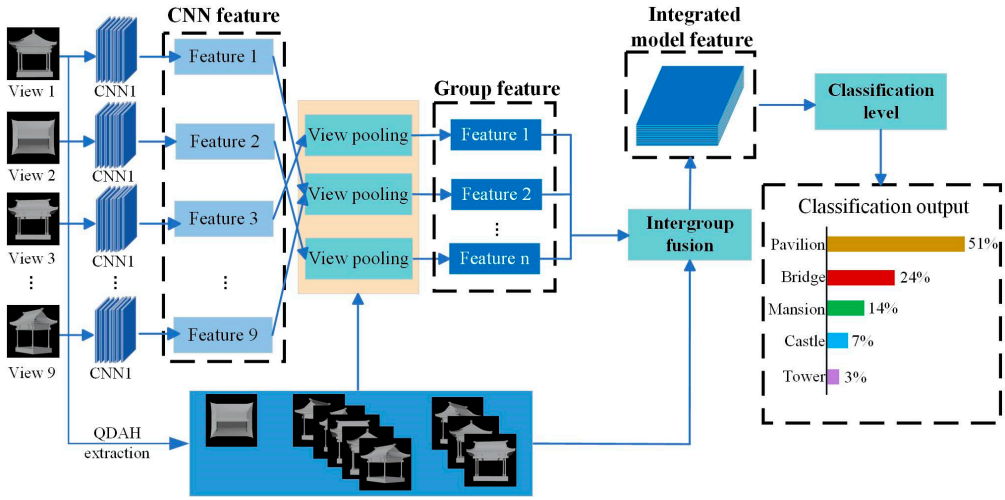


Figure 10. Network architecture of the proposed deep feature fusion model with QDAH-guided view grouping and attention-based aggregation.

Before training, this study extracts the QDAH features of each view to calculate view similarity and perform grouping. Notably, QDAH features are precomputed and used to guide view grouping, and they are kept fixed during network training; therefore, they only need to be computed once and indexed. The set of views for the building model is then input into the improved MVCNN network. The network first divides the views into groups based on the grouping results of the QDAH features, assigning the deep features of each view to the corresponding view group, with the specific grouping criteria detailed in Section 3.4.2. After grouping, the features within each group are fused, and an intergroup fusion module fuses the group features into a global feature representation, which is used to represent the 3D building model. The global features are subsequently passed to a classification layer for category prediction.

The construction of this model effectively minimizes the interference of redundant views in feature extraction through the view grouping strategy. Additionally, the attention mechanism further optimizes the importance distribution of features both within and across groups.

4. Experimental Results and Analysis

4.1. Experimental Setup

This section comprehensively evaluates the proposed retrieval framework under different compute budgets and assesses the effectiveness and robustness of the learned representations under viewpoint variations. Because our method targets perceptual and viewpoint-stable structural representation rather than geometric reconstruction accuracy, retrieval performance is adopted as an indirect measure of representation quality. We then describe the dataset construction, preprocessing, experimental environment, retrieval procedure, and evaluation metrics.

4.1.1. Dataset Collection and Preprocessing

Before extracting features, a series of preprocessing and normalization operations was applied to the 3D building models to improve data quality and ensure consistency across the dataset. Specifically, we removed model bases that are geometrically simple yet often dominant in size, as they may introduce noise and weaken feature discriminability. All models were then normalized in translation and scale by centering their geometric centroid at the origin and uniformly scaling them to a standard size, enabling fair and scale-invariant comparisons. Since most models were already upright, only a small number of misaligned cases were manually adjusted to ensure a consistent vertical orientation, avoiding the additional complexity of full rotational normalization.

Given the wide range of architectural forms, this study focuses on nine representative building categories: stadiums, mansions, bridges, towers, pavilions, castles, twin buildings, Gothic architecture, and small houses. A total of 584 3D models were collected from public repositories and benchmark datasets, including 3D Warehouse, CGTrader, Free3D, BuildingNet, and NTU. However, we did not directly adopt a single existing public benchmark for evaluation because our study focuses on 3D building retrieval driven by topological features, and the core goal is to validate how well these features capture and discriminate buildings with different topology-related structures. Although many publicly available 3D repositories and general-purpose shape databases exist, building models exhibiting rich and reliable topology-related structures is often underrepresented, and the modeling of such structures can be inconsistent across collections. This makes it difficult to conduct a controlled evaluation that directly assesses the effectiveness of the proposed topological features. Therefore, to verify the contribution of the proposed QDAH-guided topological features in this retrieval setting, we curated these models into a building-specific benchmark with consistent category definitions and sufficient topological diversity. The number of models in each category and representative samples are listed in Table 1.

Table 1. Category distribution and representative samples of the self-built 3D building model dataset.

Category	Stadium	Mansion	Tower	Pavilion	Castle	Twin Building	Gothic Architecture	House	Bridge
Quantity	84	45	52	56	45	20	125	98	56
Model sample									

Although the total number of models is 584, this curated scale is adequate for our domain-specific task and comparable to established benchmarks in the field. For example, the Engineering Shape Benchmark (ESB) contains 866 models to evaluate complex CAD structures [49]. In addition, benchmark collections designed for specific geometric or

appearance factors are frequently of comparable scale, such as a rigorous comparative benchmark built with exactly 572 textured mesh models [50], and a specialized dataset of 400 models utilized to evaluate watertight 3D shapes [51]. By rigorously covering nine representative categories with consistent category definitions and sufficient topological diversity, this curated dataset provides a robust and benchmark-aligned foundation for evaluating the proposed framework.

To ensure reproducibility for methods that require offline training, including Fusion Algorithm 2 and the deep learning baselines, we further specify the train–test split and validation strategy. The 584 models were partitioned into training and test subsets with an 80 to 20 ratio using a stratified random split that preserves the category distribution shown in Table 1. The split was generated using a fixed random seed to make the partition deterministic. During training, ten percent of the training subset was reserved as a validation set. This validation set was used to monitor convergence and to select the final checkpoint, thereby reducing the risk of overfitting. For retrieval evaluation, each model in the test subset was used as a query. The query descriptor was compared against the database of 584 models, and the query instance itself was excluded from the ranked list. All retrieval metrics were computed over the test queries only, and the test subset was not used for any form of model selection.

4.1.2. Experimental Environment

All experiments in this study were conducted on a personal computer equipped with an Intel Core i7-1065G7 processor running at 1.30 GHz with 4 cores, 8 GB of RAM, and a Windows 11 (64-bit) operating system. The entire retrieval framework was implemented using Python 3.9. Blender 3.6 was used for preprocessing and multi-view rendering, and feature extraction and retrieval were implemented using standard scientific computing libraries and deep-learning frameworks.

4.1.3. Retrieval Procedure

In this study, each 3D building model in the test subset is used as a query input to the retrieval framework. To support feature extraction, the model is first rendered into multiple 2D projection views using Blender, under consistent lighting and camera configurations. The framework is implemented as two fusion schemes (realized by Fusion Algorithm 1 and Fusion Algorithm 2) targeting different compute budgets, rather than as competing alternatives. Here, "compute budget" refers to both (i) the offline training cost and (ii) the computational cost per query at retrieval time (feature extraction and similarity matching). Fusion Algorithm 1 extracts traditional feature descriptors (e.g., Zernike moments and Fourier descriptors) from the multi-view images and combines them with topological perception features extracted using QDAH. This scheme does not require an additional learning stage and is therefore lightweight. Fusion Algorithm 2 directly uses the improved MVCNN to extract deep features from the same set of views, which are then fused with topological perception features obtained via QDAH. This scheme involves an offline training stage and incurs a higher per-query cost during retrieval due to deep-model inference for multi-view feature extraction. The system compares the query features against all models in the database and returns the top 10 most similar models based on similarity scores.

The performance of each algorithm is then evaluated using commonly adopted metrics, including overall Accuracy (Acc), First-Tier Similarity (FT), Second-Tier Similarity (ST), and precision–recall (PR) curves. Here, Acc denotes top-1 retrieval accuracy (nearest-neighbor accuracy), i.e., whether the highest-ranked result belongs to the same class as the query. Let C be the number of models in the query's class in the database (including the

query), since the query itself is excluded from the ranked list, so there are C–1 relevant models. FT is the recall within the top C–1 retrieved results, and ST is the recall within the top 2(C–1) retrieved results. PR curves are obtained by computing precision and recall at different cutoffs along the ranked retrieval list. These retrieval metrics are used to evaluate representation effectiveness in the retrieval setting; their perceptual implications are further discussed in Section 5.

4.2. Evaluation of the Matching Strategy in the Traditional-Feature Fusion Scheme

To justify the matching strategy adopted in the traditional-feature fusion scheme (Section 3.3), we compare representative-view matching with the widely used Hungarian matching, which performs full view-to-view assignment based on a complete distance matrix. Following the protocol in Section 4.1, we report three summary metrics (Acc, FT, and ST) and PR curves to characterize ranking quality across decision thresholds. This comparison is motivated by the high redundancy among building side views, where small viewpoint perturbations can yield near-equivalent matches, making full assignment potentially unnecessary.

In terms of effectiveness, the two strategies produce highly consistent results. As shown in Figure 11a, representative-view matching achieves Acc, FT, and ST comparable to Hungarian matching, with only marginal differences across metrics. The PR curves in Figure 11b are nearly overlapping, indicating that replacing full assignment with representative-view matching does not noticeably change retrieval quality under our setting.

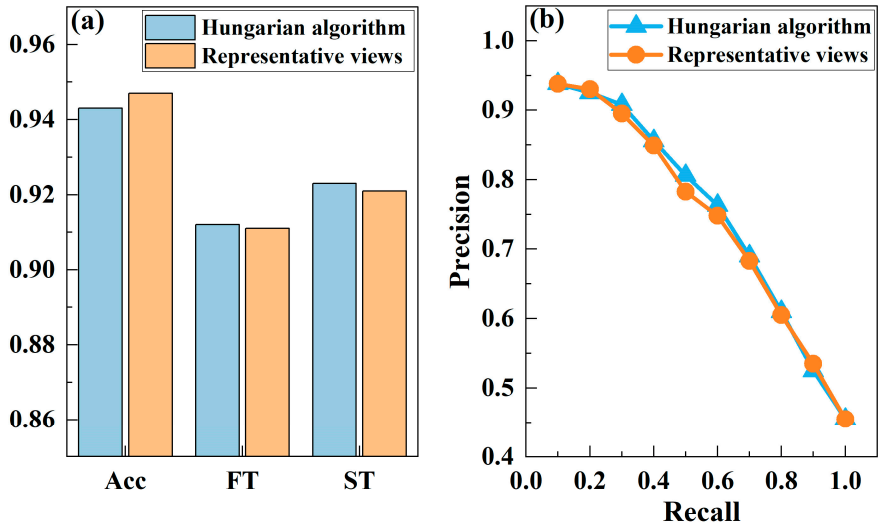


Figure 11. Comparison of traditional matching strategies: (**a**) Evaluation of Hungarian algorithm-based and Representative view-based matching strategies using Acc, FT, and ST metrics; (**b**) Performance comparison of the two matching strategies based on PR curves.

In terms of efficiency, representative-view matching is consistently faster. As reported in Table 2, its maximum, minimum, and mean running times are all lower than those of Hungarian matching. On average, representative-view matching reduces the time cost by more than one second per query, demonstrating a clear advantage for large-scale retrieval and time-sensitive applications.

Overall, these results suggest that representative-view matching offers a more favorable effectiveness–efficiency trade-off for the traditional-feature fusion scheme. We therefore adopt it as the default matching strategy in the subsequent experiments of Section 4.3, as representative-view selection removes redundant side views before matching.

Table 2. Time cost statistics (maximum, minimum, and mean) for Hungarian and Representative view-based matching strategies in 3D building model retrieval.

Matching Strategy	Maximum (s)	Minimum (s)	Mean (s)
Hungarian algorithm-based matching strategy	3.62	2.57	3.06
Representative view-based matching strategy	2.13	1.38	1.79

4.3. Qualitative Retrieval Results

This section provides qualitative retrieval examples to visualize the top-ranked retrieval results under the two fusion schemes targeting different compute budgets. Figures 12 and 13 show representative top-10 retrieval lists for multiple query models under the traditional-feature fusion scheme and the deep-feature fusion scheme, respectively. Incorrect retrievals are marked with red boxes. Quantitative comparisons against baseline methods and further analysis are provided in Section 4.4.

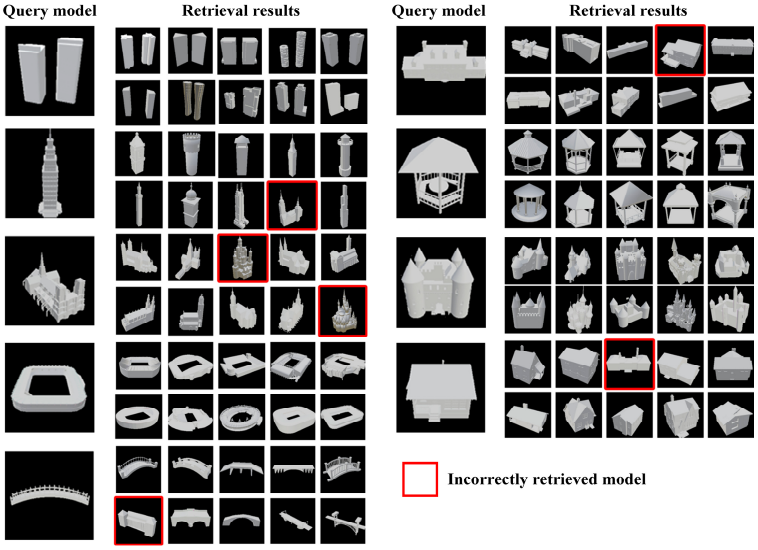


Figure 12. Top-10 retrieval results of the proposed traditional feature fusion scheme across multiple 3D building query models.

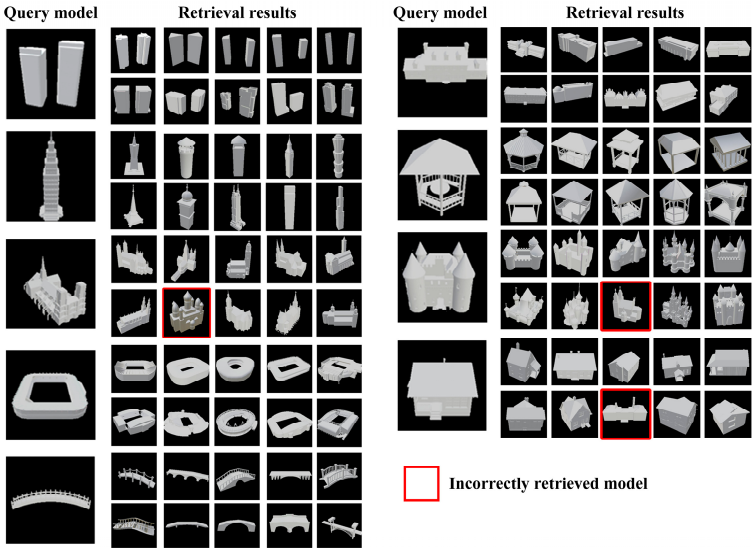


Figure 13. Top-10 retrieval results of the proposed deep feature fusion scheme across multiple 3D building query models.

4.3.1. Traditional-Feature Fusion Scheme

Figure 12 shows the qualitative results of fusing QDAH-based topological perception features with traditional feature descriptors. In the shown examples, most retrieved models are relevant to the query category, with a small number of mismatches. Specifically, six incorrect retrievals are observed in the illustrated top-10 lists, including two in the Gothic architecture category and one each in the tower, bridge, mansion, and small house categories. No incorrect retrievals are observed for the twin buildings, stadiums, and pavilions in these examples. This is consistent with the distinctive topological-perception features of these categories.

4.3.2. Deep-Feature Fusion Scheme

Figure 13 shows the qualitative results of the deep-feature fusion scheme, in which QDAH features are fused with deep features via QDAH-guided view grouping and subsequent multi-view aggregation to form a fused representation for retrieval. The retrieved lists are generally coherent across the shown queries, including categories with complex structures (e.g., Gothic architecture). Across the illustrated top-10 lists, three incorrect retrievals are observed, and these mismatched models remain visually similar to the queries in terms of overall shape. A quantitative comparison with baseline methods is reported in Section 4.4.

4.4. Comparison and Analysis

To assess the effectiveness of the proposed schemes within their respective compute settings, we compare Fusion Algorithm 1 and Fusion Algorithm 2 with representative baselines, including MVCNN [28], GVCNN [29], QDAH [19], LFD [13], and SPH [11]. Specifically, Fusion Algorithm 1 fuses QDAH with handcrafted descriptors under representative-view matching, whereas Fusion Algorithm 2 integrates QDAH with deep multi-view features via QDAH-guided view grouping and subsequent aggregation.

Figure 14 presents a qualitative comparison of hole-shaped queries. Taking the stadium model as a representative example, both Fusion Algorithm 1 and Fusion Algorithm 2 return 10/10 correct results in the top-10 list. GVCNN and QDAH achieve 9/10, followed by MVCNN (8/10), LFD (7/10), and SPH (6/10). This example suggests that incorporating QDAH into feature fusion helps preserve the salient hole-related topological perception features, thereby reducing confusion with non-hollow structures. Notably, the single error made by QDAH remains within the same QDAH-defined topological perception category as the query. In contrast, LFD and SPH produce more frequent mismatches in this example, indicating weaker discrimination for hole-shaped structures under this query setting.

To provide an overall quantitative comparison, we report Acc, ST, and FT for all methods in Table 3. Table 3 summarizes the results for the two compute settings. As the schemes target different budgets, we focus on within-setting improvements over the corresponding baselines. Fusion Algorithm 2 yields the best results within the deep-feature setting across all three metrics, improving over MVCNN and GVCNN, suggesting that QDAH-guided view grouping and aggregation help improve the robustness of multi-view deep representations. Within the traditional-feature group, Fusion Algorithm 1 consistently outperforms QDAH, LFD, and SPH, indicating that combining QDAH with complementary handcrafted descriptors offers stronger discrimination than using topological perception features alone.

Specifically, within the deep-feature group, Fusion Algorithm 2 achieves the highest across all three metrics. It improves Acc by 4.0 percentage points over MVCNN and by 1.2 percentage points over GVCNN; the corresponding gains are 6.7 and 3.2 percentage points in FT, and 6.8 and 5.6 percentage points in ST, respectively. These improvements support the

effectiveness of integrating topological-perception guidance into multi-view learning, where QDAH-guided view grouping enhances the robustness of multi-view aggregation.

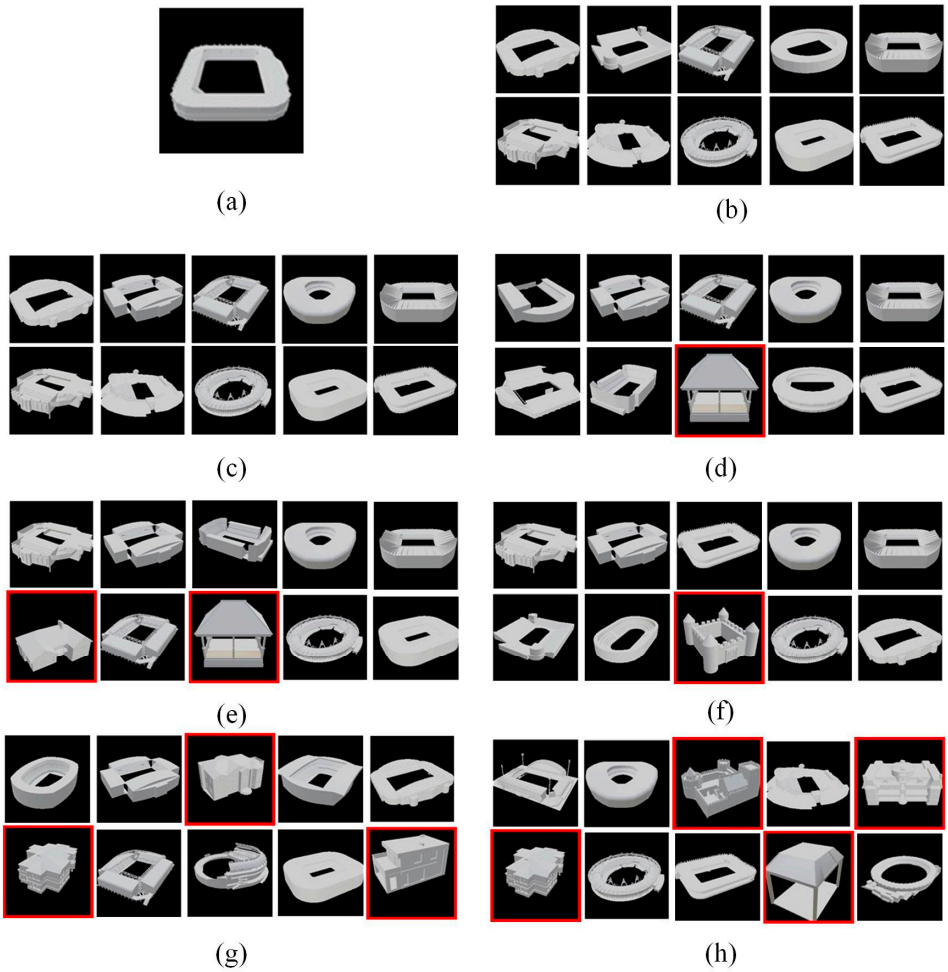


Figure 14. Qualitative comparison of 3D building model retrieval results using different algorithms: (**a**) shows the query model; (**b**–**h**) present the top 10 retrieved results of Fusion Algorithm 1, Fusion Algorithm 2, GVCNN, MVCNN, QDAH, LFD, and SPH, respectively. Incorrect retrievals are marked with red boxes.

Table 3. Overall retrieval performance of fusion and baseline algorithms across multiple evaluation metrics.

Algorithm	Algorithms Based on Traditional Features				Algorithms Based on Deep Features		
	Fusion Algorithm 1	**QDAH**	**LFD**	**SPH**	**Fusion Algorithm 2**	**MVCNN**	**GVCNN**
Acc	**0.881**	0.801	0.703	0.624	**0.963**	0.923	0.951
ST	**0.824**	0.764	0.607	0.552	**0.919**	0.851	0.863
FT	**0.866**	0.793	0.674	0.588	**0.941**	0.874	0.909

Note: Bold values indicate the best results for each evaluation metric.

Within the traditional-feature group, Fusion Algorithm 1 achieves the highest across all three metrics. Compared with QDAH, it improves Acc by 8.0 percentage points, ST by 6.0 percentage points, and FT by 7.3 percentage points. These gains indicate that combining QDAH with complementary traditional descriptors provides additional discriminative features beyond topological-perception features alone, leading to more stable retrieval performance under the traditional-feature fusion scheme. Relative to LFD and SPH, Fusion

Algorithm 1 also yields consistently higher scores, further supporting the benefit of feature fusion for traditional-feature-based retrieval.

In summary, both schemes benefit from incorporating QDAH guidance: Fusion Algorithm 1 provides an efficient option under the traditional-feature setting, while Fusion Algorithm 2 offers a higher-capacity option under the deep-feature setting. This supports our motivation that the two schemes target different compute budgets rather than competing alternatives.

To examine category-wise performance variation, we further report per-category accuracy for all methods in Table 4. Overall, within the traditional-feature group, Fusion Algorithm 1 achieves the highest accuracy across most categories and yields the best total accuracy (0.881), outperforming QDAH (0.801), LFD (0.703), and SPH (0.624). This pattern indicates that augmenting QDAH with complementary traditional descriptors improves robustness beyond using topological-perception features alone. One notable exception is the Stadium category, where QDAH slightly outperforms Fusion Algorithm 1 (0.964 vs. 0.94), suggesting that for categories with highly distinctive topological-perception signatures (e.g., hollow structures), QDAH-derived features can be highly discriminative on their own.

Table 4. Accuracy comparison of fusion methods and baseline algorithms across building categories.

Category	Algorithms Based on Traditional Features				Algorithms Based on Deep Features		
	Fusion Algorithm 1	**QDAH**	**LFD**	**SPH**	**Fusion Algorithm 2**	**MVCNN**	**GVCNN**
Stadium	0.94	**0.964**	0.869	0.845	**1**	0.892	0.904
Mansion	**0.778**	0.667	0.667	0.533	**0.956**	0.933	0.933
Bridge	**0.875**	0.785	0.714	0.661	0.928	0.911	**0.964**
Tower	**0.942**	0.865	0.826	0.692	**0.981**	0.942	0.98
Pavilion	**0.857**	0.821	0.642	0.589	**0.946**	0.928	0.911
Castle	**0.844**	0.755	0.6	0.578	**0.933**	0.888	0.888
Twin building	**0.95**	0.9	0.85	0.7	**0.95**	0.9	0.9
Gothic architecture	**0.881**	0.744	0.64	0.528	0.952	0.904	**0.984**
House	**0.867**	0.765	0.632	0.561	**0.989**	0.979	0.988
Total	**0.881**	0.801	0.703	0.624	**0.963**	0.923	0.951

Within the deep-feature group, Fusion Algorithm 2 attains the highest group-level accuracy (0.963), exceeding MVCNN (0.923) and GVCNN (0.951), and achieves the top score in most categories (e.g., Mansion, Tower, Pavilion, Castle, and House). Meanwhile, GVCNN is marginally better in Bridge and Gothic architecture (0.964 and 0.984, respectively), indicating that different deep aggregation designs may have category-dependent strengths. Overall, Table 4 indicates that incorporating QDAH guidance improves category-wise accuracy within both the traditional-feature and deep-feature settings, while the relative ranking among methods can vary across categories.

To rigorously verify that the performance gains are robust and not subject to random variance across different architectural types, a paired *t*-test was conducted. Because FT and ST are evaluated as overall metrics across the entire dataset (Table 3), this statistical analysis specifically utilizes the category-wise accuracy (Acc) values detailed in Table 4 to assess cross-category consistency. Table 5 presents these test results, comparing the proposed fusion algorithms against their respective primary baselines (i.e., Fusion Algorithm 1 vs. QDAH, Fusion Algorithm 2 vs. MVCNN, and Fusion Algorithm 2 vs. GVCNN).

Table 5. Category-level statistical significance test (paired *t*-test) comparing the proposed fusion algorithms against their primary baselines across nine building categories. ** indicates statistical significance at the $p < 0.01$ level.

Comparison	Mean Difference (ΔAcc)	*t*-Statistic	*p*-Value (Two-Sided)
Fusion Algorithm 1 vs. QDAH	0.074	4.637	0.0016 **
Fusion Algorithm 2 vs. MVCNN	0.040	3.999	0.0038 **
Fusion Algorithm 2 vs. GVCNN	0.020	1.430	0.1842

As shown in Table 5, the comparisons of Fusion Algorithm 1 vs. QDAH and Fusion Algorithm 2 vs. MVCNN yield *p*-values of 0.0016 and 0.0038, respectively. Both are well below the strict significance threshold of 0.01 ($p < 0.01$). This demonstrates that the accuracy improvements achieved by incorporating QDAH-guided topological perception are statistically highly significant, underscoring the robustness of the observed gains across diverse building categories in our benchmark.

When comparing Fusion Algorithm 2 with GVCNN, the mean accuracy difference is 0.020 with a *p*-value of 0.1842 ($p > 0.05$). This suggests that although Fusion Algorithm 2 achieves a higher overall accuracy (0.963 vs. 0.951, as shown in Table 3), the category-level accuracy difference is not statistically significant under the category-wise paired test ($t(8) = 1.430$, $p = 0.1842$), implying comparable performance between the two methods at this aggregation level. A plausible reason is that GVCNN is already a highly competitive baseline with an effective multi-view strategy, leading to performance saturation that limits the potential for statistically significant gains in category-level accuracy. Nevertheless, as shown by the PR curves in Figure 15, the QDAH-guided view grouping in Fusion Algorithm 2 provides a more robust ranking performance across most recall levels. Overall, the paired *t*-test results support that the proposed framework provides statistically reliable improvements over its corresponding baselines, where significance is observed and demonstrates stable performance across diverse architectural categories in our benchmark.

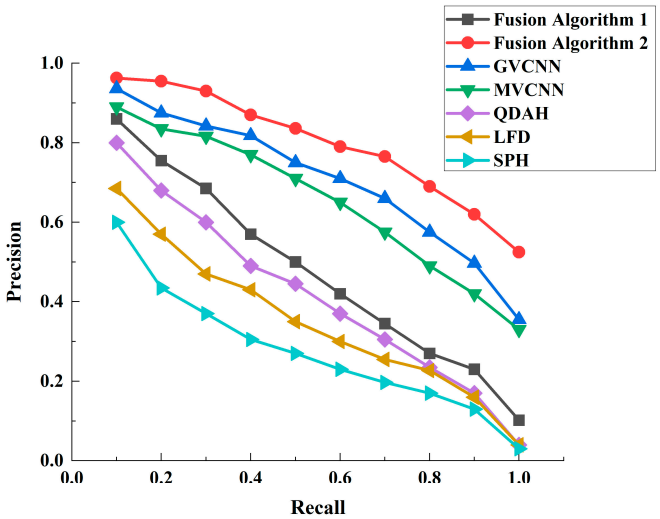


Figure 15. Performance comparison of fusion and baseline algorithms based on PR curves.

To further validate the retrieval performance, we plot the PR curves for all methods, as shown in Figure 15. Within the deep-feature setting, Fusion Algorithm 2 achieves higher precision than the evaluated deep-feature baselines at most recall levels, suggesting improved ranking quality within this budget. Methodologically, Fusion Algorithm 2 performs QDAH-guided view grouping prior to multi-view aggregation, which can mitigate the

influence of redundant views and improve robustness to viewpoint variation. This design is reflected in the PR curves, where Fusion Algorithm 2 maintains a consistently higher precision trajectory across most recall levels.

Under the traditional-feature setting, traditional handcrafted baselines generally exhibit lower PR performance than deep-feature methods, which is expected given differences in representation capacity and the available compute budget. Nevertheless, fusing QDAH into traditional handcrafted descriptors can improve ranking quality within this setting. Specifically, Fusion Algorithm 1 yields better PR performance than other traditional-feature methods, indicating that topological-perception features complement geometry-only descriptors and enhance ranking quality.

5. Discussion

The experimental results show that deep-feature-based methods achieve higher absolute accuracy across building model categories, which is expected given their stronger representation capacity under higher compute budgets. We therefore examine the marginal contribution of incorporating topological-perception features through feature fusion within each compute setting. As reported in Table 3, Fusion Algorithm 1 reaches a total accuracy of 0.881 and substantially improves over traditional-feature baselines, including QDAH (0.801), LFD (0.703), and SPH (0.624). Fusion Algorithm 2 attains the highest overall accuracy (0.963) and further improves upon deep-feature baselines such as MVCNN (0.923) and GVCNN (0.951), although the absolute gain is smaller than that observed in the traditional-feature setting. This observation aligns with prior 3D shape retrieval literature, where deep multi-view representations generally outperform handcrafted descriptors, and additional improvements beyond strong deep baselines are often modest [28,29].

This difference in marginal gain is consistent with the complementary roles that topological-perception features play in the two schemes. In the traditional-feature fusion scheme, traditional handcrafted descriptors tend to be sensitive to view sampling and often under-represent global structural features; integrating QDAH features introduces viewpoint-stable, perceptually salient topological-perception information, which compensates for these weaknesses and yields a larger gain. In the deep-feature fusion scheme, multi-layer convolutional representations already encode part of the global structure through hierarchical abstraction, so the benefit of QDAH-based topological-perception features is mainly reflected in improving view organization and aggregation robustness (via QDAH-guided view grouping), leading to a more moderate but consistent improvement.

Beyond retrieval performance, the proposed topological-perception features hold potential for 3D model reconstruction applications at a structural and perceptual level. They can serve as perceptual priors by emphasizing salient holes and concavities that should be preserved in building shapes, and can further provide high-level structural constraints or guidance by retrieving reference models with similar topological-perception patterns to guide modeling or editing. In addition, these features can be used as consistency checks by assessing whether a modeled or reconstructed building shape preserves hole-and-concavity patterns consistent with the retrieved structural references. Nevertheless, their role is auxiliary, and they are not sufficient to replace geometry-based or learning-based representations when fine-grained shape details are required.

From a practical deployment standpoint, the proposed framework is also scalable in terms of computational cost and applicability to larger databases. For the traditional-feature fusion scheme, Fusion Algorithm 1 combines lightweight descriptors with representative-view matching; as reported in Table 2, this reduces the mean matching time per query from 3.06 s (Hungarian matching) to 1.79 s, improving efficiency for retrieval settings where matching overhead is critical. For the deep-feature fusion scheme, Fusion Algorithm 2

requires deep feature extraction, while the QDAH descriptors used for view grouping can be computed offline and reused. With a fixed number of projected views, online retrieval reduces to descriptor similarity search. For larger repositories, the online retrieval time mainly increases with the number of database models, because each query needs to be compared against precomputed feature representations in the database. To facilitate massive-scale retrieval, similarity search can be further accelerated using standard indexing techniques (e.g., approximate nearest-neighbor search) when scaling to much larger databases. Sensitivity analyses also indicate that the framework is not overly sensitive to moderate parameter variations: Figure 6 shows that retrieval precision peaks at 0.6 and remains stable within 0.5–0.7, and Figure 9 shows that accuracy peaks at three view groups before gradually decreasing as grouping becomes overly fine-grained.

Despite the overall improvements, retrieval failures still occur. In the qualitative Top-10 examples, Figure 12 highlights 6 incorrect retrievals for the traditional fusion scheme, whereas Figure 13 shows 3 for the deep fusion scheme. Most failures arise when different buildings share highly similar global topologies, while their differences are mainly concentrated in subtle local geometric details. Under certain viewpoints, a closed hole and a deep but unclosed concavity can yield highly similar projections, so some ambiguity may persist even with angle refinement. This is consistent with Table 5, where Fusion Algorithm 2 shows no statistically significant category-level advantage over the strong deep baseline GVCNN (p = 0.1842), suggesting that remaining errors are concentrated in inherently ambiguous structures that are difficult to separate under projection-based retrieval.

Aside from these specific failure cases, several broader limitations must be acknowledged. First, this study does not evaluate the proposed method on general-purpose 3D shape benchmarks. Although curated from publicly available sources, our dataset is a building-specific collection rather than a widely adopted standard benchmark dataset. As a result, the evaluation may not fully reflect the diversity of topologically complex architectural structures found in broader publicly available 3D datasets and benchmarks, which limits cross-study comparability. More broadly, the absence of large-scale, high-quality 3D building benchmarks—with sufficient category diversity, intra-class variability, and consistent annotations—constrains both model development and fair comparison. Second, fusion remains constrained by cross-feature alignment. QDAH delivers topological-perception-oriented features, while traditional and deep descriptors represent geometry or appearance in a continuous embedding space. When these feature spaces are not well aligned, the contribution from QDAH may become less stable, limiting the overall benefit of fusion. Under complex building geometry and viewpoint-induced projection changes, this mismatch can lead to unstable grouping decisions or redundant aggregated representations, which may restrict further improvements. Third, topological perception features such as holes and concavities are inherently non-unique across buildings and are mainly summarized from projected views; therefore, they cannot be interpreted as unique identifiers or complete descriptors for detailed building reconstruction.

Finally, in addition to the above limitations, although topological perception features are motivated by perceptual principles of viewpoint invariance and global priority, our evaluation primarily relies on retrieval metrics, including Acc, ST, FT, and PR curves. These metrics provide an objective and reproducible surrogate measure for representation effectiveness and robustness in the retrieval setting, but they do not directly incorporate human perceptual judgments as an explicit evaluation criterion. Therefore, the current analysis cannot fully confirm perceptual consistency from the perspective of human subjective assessment, and complementary user-centered evaluations are needed to more directly assess perceptual consistency under viewpoint variations.

6. Conclusions

3D building model retrieval is a key module in real-scene 3D modeling, geographic information systems, and smart city development. Topological-perception features, characterized by global invariance and perceptual saliency, are well suited for capturing salient structural properties of building models, particularly under complex geometries and high intra-class variability. In this study, we develop a retrieval framework with two complementary fusion schemes targeting different compute budgets: Fusion Algorithm 1 fuses topological-perception features with handcrafted geometric descriptors, whereas Fusion Algorithm 2 fuses topological-perception features with deep multi-view representations.

The experimental results indicate that fusing topological-perception features benefits both compute settings. Under the traditional-feature setting, Fusion Algorithm 1 adopts representative-view matching to select a compact yet informative set of views, reducing redundancy and computational cost while preserving key structural information for retrieval. Under the deep-feature setting, Fusion Algorithm 2 performs QDAH-guided view grouping prior to multi-view aggregation, which improves view consistency and enhances the robustness of the aggregated representation to viewpoint variation. Together, these scheme-specific designs provide practical retrieval options under different resource constraints while consistently leveraging topological-perception information.

Future work will focus on constructing larger and more diverse 3D building benchmarks with consistent annotations to enable fairer evaluation. To examine generalization beyond our current building-specific collection, experiments will also be extended to additional publicly available datasets and benchmarks, particularly those exhibiting richer topological variations. We will then explore more effective fusion and alignment mechanisms between topological-perception features and geometry or appearance-driven representations, so that the complementary information from different feature spaces can be integrated more consistently. In addition, we plan to investigate alternative projection and view-sampling strategies, such as overlapping or adaptive sampling, to improve cross-view continuity and further enhance robustness in complex scenes. We will further conduct user-centered evaluations to directly assess perceptual consistency under viewpoint variations and to complement retrieval-based metrics.

Author Contributions: Conceptualization, Xinwen Zhang and Yuan Ding; methodology, Xinwen Zhang; software, Xinwen Zhang; validation, Yi Lu, Jin Zhu, Xiaoping Rui and Hua Shao; investigation, Xiaoping Rui and Yi Lu; resources, Xinwen Zhang and Xiaoping Rui; data curation, Xinwen Zhang; writing—original draft preparation, Xinwen Zhang; writing—review and editing, Xinwen Zhang, Yuan Ding, Hua Shao, Jin Zhu and Xiaoping Rui; visualization, Xinwen Zhang; supervision, Yuan Ding; project administration, Yuan Ding and Yi Lu; funding acquisition, Yuan Ding. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China (Grant Nos. 42471488, 42371397, and 42301474), the Fundamental Research Funds for the Central Universities (Grant No. B250201132), and the Marine Science and Technology Innovation Project of Jiangsu Province (Grant No. JSZRHYKJ202302).

Data Availability Statement: The 3D model data used in this study were obtained from third-party sources and are subject to copyright restrictions that prevent their public distribution. The custom code developed for analysis is dependent on these proprietary datasets and is not independently functional. Therefore, neither the original datasets nor the analysis code can be made publicly available. Detailed methodologies are described within the article to facilitate reproducibility. For further inquiries, interested researchers are encouraged to contact the authors.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Chen, R. The development of 3D city model and its applications in urban planning. In *2011 19th International Conference on Geoinformatics*; IEEE: Piscataway, NJ, USA, 2011.
2. Gimenez, L.; Hippolyte, J.-L.; Robert, S.; Suard, F.; Zreik, K. Reconstruction of 3D building information models from 2D scanned plans. *J. Build. Eng.* **2015**, *2*, 24–35. [CrossRef]
3. Tang, Y.M.; Ho, H.L. 3D modeling and computer graphics in virtual reality. In *Mixed Reality and Three-Dimensional Computer Graphics*; IntechOpen: London, UK, 2020.
4. Croce, V.; Caroti, G.; Piemonte, A.; Bevilacqua, M.G. Geomatics for Cultural Heritage conservation: Integrated survey and 3D modeling. In Proceedings of the IMEKO TC4 International Conference on Metrology for Archaeology and Cultural Heritage, MetroArchaeo, Florence, Italy, 19–21 October 2019.
5. Ning, F.; Shi, Y.; Tong, X.; Cai, M.; Xu, W. A review and assessment of 3D CAD model retrieval in machine-part design. *Int. J. Comput. Integr. Manuf.* **2025**, *38*, 752–774. [CrossRef]
6. Tsai, Y.-Z.; Luo, J.; Wang, Y.; Chen, J. The World is Too Big to Download: 3D Model Retrieval for World-Scale Augmented Reality. In *Proceedings of the 14th Conference on ACM Multimedia Systems*; ACM: New York, NY, USA, 2023.
7. Keshavarzi, M.; Hotson, C.; Cheng, C.-Y.; Nourbakhsh, M.; Bergin, M.; Asl, M.R. Sketchopt: Sketch-based parametric model retrieval for generative design. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*; ACM: New York, NY, USA, 2021.
8. Yang, Y.; Lin, H.; Zhang, Y. Content-based 3-D model retrieval: A survey. *IEEE Trans. Syst. Man Cybern. Part C* **2007**, *37*, 1081–1098. [CrossRef]
9. Li, P.; Ma, H.; Ming, A. 3D model retrieval using 2D view and transform-based features. In *Proceedings of the Advances in Multimedia Information Processing-PCM 2010: 11th Pacific Rim Conference on Multimedia, Shanghai, China, 21–24 September 2010*; Part I 11; Springer: Berlin/Heidelberg, Germany, 2010.
10. Pan, X.; Zhang, Y.; Zhang, S.; Ye, X. Radius-normal histogram and hybrid strategy for 3D shape retrieval. In *International Conference on Shape Modeling and Applications 2005 (SMI'05)*; IEEE: Piscataway, NJ, USA, 2005.
11. Kazhdan, M.; Funkhouser, T.; Rusinkiewicz, S. Rotation invariant spherical harmonic representation of 3d shape descriptors. In *Symposium on Geometry Processing*; Eurographics Association: Goslar, Germany, 2003.
12. Bandlow, T.; Hauck, A.; Einsele, T.; Farber, G. Recognising objects by their silhouette. In *IMACS Conference on Computational Engineering in Systems Applications (CESA'98)*; IEEE: Piscataway, NJ, USA, 1998.
13. Chen, D.Y.; Tian, X.-P.; Shen, Y.-T.; Ouhyoung, M. On visual similarity based 3D model retrieval. *Comput. Graph. Forum* **2003**, *22*, 223–232. [CrossRef]
14. Guo, Y.; Bennamoun, M.; Sohel, F.; Lu, M.; Wan, J.; Kwok, N.M. A comprehensive performance evaluation of 3D local feature descriptors. *Int. J. Comput. Vis.* **2016**, *116*, 66–89. [CrossRef]
15. Li, B.; Johan, H. 3D model retrieval using hybrid features and class information. *Multimed. Tools Appl.* **2013**, *62*, 821–846. [CrossRef]
16. Gordo, A.; Almazán, J.; Revaud, J.; Larlus, D. Deep image retrieval: Learning global representations for image search. In *Computer Vision–ECCV 2016: 14th European Conference*; Springer: Amsterdam, The Netherlands, 2016.
17. Esteves, C.; Xu, Y.; Allec-Blanchette, C.; Daniilidis, K. Equivariant multi-view networks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2019.
18. Wei, N.; Zhou, T.; Zhang, Z.; Zhuo, Y.; Chen, L. Visual working memory representation as a topologically defined perceptual object. *J. Vis.* **2019**, *19*, 12. [CrossRef] [PubMed]
19. Li, C.; Ding, Y.; Yang, Y. A Novel Method Based on Topological Perception Theory for 3D Landmark Building Model Retrieval. *Appl. Sci.* **2024**, *14*, 1146. [CrossRef]
20. Ohbuchi, R.; Osada, K.; Furuya, T.; Banno, T. Salient local visual features for shape-based 3D model retrieval. In *2008 IEEE International Conference on Shape Modeling and Applications*; IEEE: Piscataway, NJ, USA, 2008.
21. Tu, H. A 3D model retrieval method based on multi-feature fusion. *Int. J. Inf. Commun. Technol.* **2019**, *15*, 121–131. [CrossRef]
22. Mahmoudi, S.; Daoudi, M. 3D models retrieval by using characteristic views. In *2002 International Conference on Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2002.
23. Dubey, S.R. A decade survey of content based image retrieval using deep learning. *IEEE Trans. Circuits Syst. Video Technol.* **2021**, *32*, 2687–2704. [CrossRef]
24. Obaid, K.B.; Zeebaree, S.; Ahmed, O.M. Deep learning models based on image classification: A review. *Int. J. Sci. Bus.* **2020**, *4*, 75–81. [CrossRef]
25. Ye, F.; Dong, M.; Luo, W.; Xiao, H.; Zhao, X.; Min, W. Agricultural remote sensing image retrieval based on convolutional neural network and reranking. *Trans. Chin. Soc. Agric. Eng.* **2019**, *35*, 138–145. [CrossRef]
26. Wang, Y.; Lin, X.; Wu, L.; Zhang, W. Effective multi-query expansions: Collaborative deep networks for robust landmark retrieval. *IEEE Trans. Image Process.* **2017**, *26*, 1393–1404. [CrossRef]

27. Gao, L.; Wu, Y.; Yang, T.; Zhang, X.; Zeng, Z.; Chan, C.K.D.; Chen, W. Research on Image Classification and Retrieval Using Deep Learning with Attention Mechanism on Diaspora Chinese Architectural Heritage in Jiangmen, China. *Buildings* **2023**, *13*, 275. [CrossRef]
28. Su, H.; Maji, S.; Kalogerakis, E.; Learned-Miller, E. Multi-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE International Conference on Computer Vision*; IEEE: Piscataway, NJ, USA, 2015.
29. Feng, Y.; Zhang, Z.; Zhao, X.; Ji, R.; Gao, Y. Gvcnn: Group-view convolutional neural networks for 3d shape recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*; IEEE: Piscataway, NJ, USA, 2018.
30. Sun, K.; Zhang, J.; Liu, J.; Yu, R.; Song, Z. DRCNN: Dynamic routing convolutional neural network for multi-view 3D object recognition. *IEEE Trans. Image Process.* **2020**, *30*, 868–877. [CrossRef]
31. Chen, L. Topological structure in visual perception. *Science* **1982**, *218*, 699–700. [CrossRef] [PubMed]
32. Chen, L. The topological approach to perceptual organization. *Vis. Cogn.* **2005**, *12*, 553–637. [CrossRef]
33. Tozzi, A.; Peters, J.F. Topology of human perception. *Prepr. Res. Gate* **2019**. [CrossRef]
34. Peng, P.; Yang, K.-F.; Luo, F.-Y.; Li, Y.-J. Saliency detection inspired by topological perception theory. *Int. J. Comput. Vis.* **2021**, *129*, 2352–2374. [CrossRef]
35. Fei, F.; Nagai, Y. Topological perception on attention to product shape. *Int. J. Des. Creat.* **2020**, *8*, 88–101. [CrossRef]
36. Ti, P.; Xiong, T.; Qiu, Y.; Wang, L.; Li, Z. Automatic generation of outline-based representations of landmark buildings with distinctive shapes. *Int. J. Geogr. Inf. Sci.* **2023**, *37*, 864–884.
37. Elias, B.; Paelke, V. User-centered design of landmark visualizations. In *Map-Based Mobile Services: Design, Interaction and Usability*; Springer: Berlin/Heidelberg, Germany, 2008; pp. 33–56.
38. Bardinet, E.; Fernández-Vidal, S.; Damas, S.; Malandain, G.; de la Blanca Capilla, N.P. Structural object matching. In Proceedings of the 2nd International Symposium on Advanced Concepts for Intelligent Vision Systems (ACIVS 2000), Baden-Baden, Germany, 3–4 August 2000.
39. Hilaga, M.; Shinagawa, Y.; Kohmura, T.; Kunii, T.L. Topology matching for fully automatic similarity estimation of 3D shapes. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*; ACM: New York, NY, USA, 2001.
40. Okuneva, G.; Kunii, T.; Shinagawa, Y. *Topology of Integrable Variants of Nonholonomic Suslov's Problem*; Technical Report 94-1-033; Department of Computer Software, The University of Aizu: Aizuwakamatsu, Japan, 1994.
41. Sun, Y.; Li, F.; Li, H.; Song, Y.; Wang, W.; Zhou, R.; Xiong, J.; He, W.; Peng, Y.; Liu, Y.; et al. Performance of Topological Perception in the Myopic Population. *Curr. Eye Res.* **2020**, *45*, 1458–1465. [CrossRef]
42. Xie, J.; Dai, G.; Fang, Y. Deep multimetric learning for shape-based 3D model retrieval. *IEEE Trans. Multimed.* **2017**, *19*, 2463–2474. [CrossRef]
43. Huang, Y.; Huang, K.; Tan, T.; Tao, D. A novel visual organization based on topological perception. In *Computer Vision–ACCV 2009: 9th Asian Conference on Computer Vision*; Springer: Berlin/Heidelberg, Germany, 2010.
44. Furukawa, Y.; Hernández, C. *Multi-View Stereo: A Tutorial, Foundations and Trends® in Computer Graphics and Vision*; Now Publishers Inc.: Hanover, Germany, 2015.
45. Li, B.; Mao, B. Similarity Calculation of 3D Building Model based on EMD of Graph Features. In *2023 26th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*; IEEE: Piscataway, NJ, USA, 2023.
46. Sokic, E.; Konjicija, S. Phase preserving Fourier descriptor for shape-based image retrieval. *Signal Process. Image Commun.* **2016**, *40*, 82–96. [CrossRef]
47. Novotni, M.; Klein, R. 3D Zernike descriptors for content based shape retrieval. In *Proceedings of the Eighth ACM Symposium on Solid Modeling and Applications*; ACM: New York, NY, USA, 2003.
48. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*; IEEE: Piscataway, NJ, USA, 2005.
49. Jayanti, S.; Kalyanaraman, Y.; Iyer, N.; Ramani, K. Developing an engineering shape benchmark for CAD models. *Comput.-Aided Des.* **2006**, *38*, 939–953.
50. Biasotti, S.; Cerri, A.; Aono, M.; Hamza, A.B.; Garro, V.; Giachetti, A.; Giorgi, D.; Godil, A.; Li, C.; Sanada, C.; et al. Retrieval and classification methods for textured 3D models: A comparative study. *Vis. Comput.* **2016**, *32*, 217–241.
51. Giorgi, D.; Biasotti, S.; Paraboschi, L. SHape REtrieval Contest 2007: Watertight Models Track. In *SHREC 2007: 3D Shape Retrieval Contest*; Veltkamp, R.C., ter Haar, F.B., Eds.; Technical Report UU-CS-2007-015; Utrecht University: Utrecht, The Netherlands, 2007.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.