Descriptor Impact on Multimodal 3D Retrieval

Maria-Eirini Pegia
CERTH-ITI / Reykjavík University
Thessaloniki, Greece / Reykjavík, Iceland
mpegia@iti.gr

Björn Þór Jónsson
Reykjavik University
Reykjavík, Iceland
bjorn@ru.is

Anastasia Moumtzidou
CERTH-ITI
Thessaloniki, Greece
moumtzid@iti.gr

Sotiris Diplaris
CERTH-ITI
Thessaloniki, Greece
diplaris@iti.gr

Ilias Gialampoukidis
CERTH-ITI
Thessaloniki, Greece
heliasgj@iti.gr

Stefanos Vrochidis
CERTH-ITI
Thessaloniki, Greece
stefanos@iti.gr

Ioannis Kompatsiaris
CERTH-ITI
Thessaloniki, Greece
ikom@iti.gr

***Abstract*—With the evolution of 3D tools, there is now plenty of 3D data for digital applications. This includes 3D retrieval, which seeks to manage such data across varied representations like point clouds, meshes, and multi-view images. However, efficiently utilizing these representations for retrieval poses a challenge. This paper evaluates different representations of each modality in uni-modal retrieval and explores optimal combinations for multi-modal retrieval. Results indicate MuseHash's superiority in MAP metric, while CMCL excels in recall. This study expands existing research by providing insights into optimal representations and combinations for 3D retrieval.**

***Keywords*—Information Retrieval, Multimedia Databases, Supervised Learning**

I. Introduction

Recent developments in 3D modeling tools [1], scanning technology [2], and consumer devices embedding 3D sensors [3] have greatly increased the accessibility of vast quantities of 3D content. This evolution has had significant impact across multiple sectors, including entertainment, gaming, healthcare [4], archaeology [5], computer-aided design (CAD) [6], and autonomous systems [7]. These advancements not only simplify 3D content creation for individuals but also revolutionize industries by streamlining design processes and facilitating more informed decision-making [8]–[12].

The community has explored various representations for 3D data, including point clouds, meshes, and multi-view images; see Figure 1. In *point clouds* [13], [14], each point corresponds to a position in space, encapsulating spatial information within the mode, while *meshes* [15], [16] are constructed from interconnected triangles to approximate the surface shape of objects. Finally, *multi-view images* [17], [18] have emerged as an effective representation, comprising a series of images captured from different viewpoints of 3D shapes.

Efficient 3D model retrieval faces a key challenge in the structured representation of data [19]. This challenge is to develop methods that can efficiently utilise these diverse representations to ensure accurate and fast retrieval. Point clouds offer detail but demand computational resources, meshes provide structure but require careful processing, and multi-view images offer comprehensive visuals but may necessitate significant storage and processing.

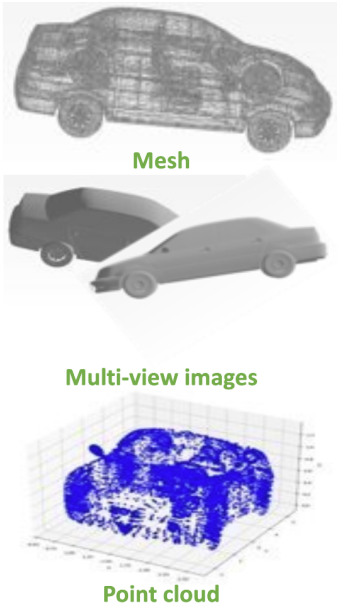

Fig. 1. Examples of 3D representations.

Recent studies highlight the significance of utilising diverse 3D representations, by considering them as distinct modalities of the same 3D objects and integrating various modalities to improve retrieval performance [20]–[22]. We have recently introduced MuseHash [22], a supervised method adapted from the multimodal image domain to 3D retrieval. The results of [22] indicated that (a) MuseHash outperformed CMCL, the state-of-the-art algorithm from the 3D domain, and (b) a combination of mesh and multi-view image modalities gave the best retrieval performance.

This paper is intended to address three shortcomings of the original work. First, due to space limitations, we only considered one representation of each modality. Second, again due to space limitations, the performance of the point cloud and multi-view images were not studies in a uni-modal scenario; since multi-view images contributed to the best retrieval performance, it is important to understand their uni-modal per-

formance. Second, we were not aware of the CMIC algorithm, a recent publicly available 3D retrieval method [21], which also outperformed the CMCL algorithm. In this paper, we thus perform a complete study of all three 3D modalities, both as uni-modal representations and as multimodal representations, comparing MuseHash to the two main competitors from the 3D domain, namely the CMCL and CMIC algorithms, aiming to identify the most beneficial combinations for 3D retrieval.

This paper thus expands the existing research [22], making the following main contributions:

- We compare MuseHash with the state-of-the-art methods from the 3D retrieval domain, to evaluate its performance against recent approaches. MuseHash outperforms all methods based on the importance of MAP metric, while CMCL works well with the recall metric.
- We compare two representations of all three modalities—point clouds, meshes, and multi-view images—in a uni-modal retrieval scenario, to determine the optimal representation of each modality. We find that GoogleNet, MeshNet and DGCNN are the best representations of the image, mesh and point cloud modalities, respectively.
- We comprehensively explore all multimodal combinations, to identify the optimal combinations for 3D retrieval. As before, we find that the combination of the mesh and image modalities yields the best results.

The remainder of this paper is organised as follows: Section II offers an overview of related literature, while Section III delves into the detailed description of the proposed evaluation. Subsequently, Section IV showcases the experimental findings, leading to a summary in Section V to conclude the paper.

II. Related work

Although there are several representations of 3D data, the bulk of the research work focused on images, meshes and point clouds. Therefore we will limit our research only to these methods. In the domain of 3D retrieval there is a plethora of methods, distinguished into uni-modal and multimodal methods based on the number of the modalities that they use. Uni-modal methods use only one modality (e.g., mesh or point cloud) while multimodal methods two or more modalities (e.g., mesh and point cloud).

For the mesh data, MeshNet [15] and MeshCNN [16] represent two cutting-edge approaches that exclusively utilise mesh data. MeshNet operates by converting mesh data into sequences of faces, where each face is associated with spatial and structural vectors, which are then combined using a multi-layer perceptron. Conversely, MeshCNN performs convolution and pooling operations directly on mesh edges and neighboring triangle edges, thus maintaining the inherent structure of the mesh throughout pooling.

For the point cloud data, PointNet [14], for instance, is an architecture designed to effectively handle unordered point clouds, offering a holistic end-to-end solution for tasks such as classification and retrieval. DGCNN [13], on the other hand, employs dynamic graph convolution for processing point clouds, although challenges persist due to the sparsity and irregularity inherent in point cloud data.

For the image data, VGG16 [23], AlexNet [24], ResNet50 [25] and GoogleNet [26] are popular deep learning architectures used for image retrieval. Based on a recent survey [27], ResNet50 and GoogleNet stand out as reliable options for extracting visual features, because they are more stable than the others across different 3D datasets.

Turning these approaches into 3D multimodal retrieval, other approaches leverage diverse representations to conduct combined queries, enhancing retrieval performance by exploiting multiple sources of information. Cross-Modal Center Loss (CMCL) [20] integrates point clouds, meshes, and multi-view images into a unified framework. This approach collectively trains representations from various 3D modalities to identify optimal features. Various loss functions, such as cross-entropy and mean-square-error, are utilised to refine and enhance the framework's performance. Additionally, Cross Modal Instance-Category (CMIC) [21] is a technique for image based 3D shape retrieval that utilises cross-modal contrastive learning from instance and category levels. It introduces a color transfer mechanism as a powerful data augmentation technique for improving contrastive learning.

In addition to the mentioned methods, a recent work [22] extends techniques from image retrieval to 3D retrieval. For instance, Label-Attended Hashing (LAH) [28] initially generates embedding for images and label co-occurrence separately, then integrates them using a graph convolutional network to boost the model's capabilities. Similarly, Supervised Bayesian Hashing for Multimodal Image Representation (MuseHash) [29] estimates semantic probabilities and statistical properties during retrieval, showing promise in multimodal retrieval and aligning well with the complexities of 3D data.

In this study we will evaluate all models in the uni-modal scenario. Additionally, we will explore the utilisation of various aforementioned techniques in the multimodal scenario to assess which descriptors enhance the performance of multimodal approaches.

III. Methodology

To formally address the problem, we establish the following scenario: Let $\mathcal{Q}$ represent a query object and $\mathcal{DB}$ denote a database containing a collection of 3D objects represented in various views, including images and meshes. The primary goal is to conduct efficient retrieval, aiming to identify objects in $\mathcal{DB}$ that exhibit similarities with $\mathcal{Q}$. This retrieval process entails a detailed analysis of the unique features characterizing $\mathcal{Q}$, followed by a comparison of these features with corresponding attributes of objects in $\mathcal{DB}$ to determine relevant matches.

The framework consists of three distinct phases: training, offline, and querying. During the training phase, data is input into a designated architecture, resulting in the generation of feature vectors. In the offline phase, features are extracted from a retrieval set and stored in a database for future reference. In the online phase, the architecture is applied to queries, and relevant results are retrieved from the database.

In 3D retrieval, ResNet50 and GoogleNet are uni-modal image based techniques, while MeshNet and MeshCNN are uni-modal mesh based methods. Similarly, PointNet and DGCNN focus on point clouds within uni-modal approaches. In contrast, CMCL and CMIC are cross-modal 3D retrieval methods, aiming to unify features derived from various sources.

Apart from these approaches, we will utilise LAH and MuseHash, originally from the image retrieval domain but adapted for 3D retrieval [22]. LAH learns hash codes using a non-linear hash function applied to the features, while MuseHash employs Bayesian ridge regression to learn hash functions, enabling both uni-modal and multimodal queries by leveraging the same models to extract features from all modalities.

Therefore, our study covers image and point cloud modalities in uni-modal scenarios and explores how various descriptors impact retrieval performance in multimodal approaches. We aim to determine the importance of descriptor selection and find optimal combinations for improved retrieval results.

IV. Experiments

In this section, we first discuss experimental setup (Section IV-A). Then, we present results for individual modalities (Section IV-B). Next, we showcase multimodal results (Section IV-C), using the best representation of each modality. Finally, we offer some conclusive insights (Section IV-D).

A. Experimental Setup

We consider the following two datasets, commonly used in the 3D literature:

BuildingNet_v0 The BuildingNet_v0 [30] (BNv0 for short) offers extensive annotations and encompasses diverse building types such as churches, hotels, and more.

ModelNet40 The ModelNet40 [31] (MN40 for short) is a vast collection of 3D CAD models, spanning various object categories such as airplanes, offices, and more.

We investigate the influence of hash code length (16, 32, 64, 128) on several evaluation metrics, including mean average precision (MAP), precision at k (precision@k), recall at k (recall@k), and fscore at k (fscore@k), where k ranges within $\{10, 25, 50\}$, within image retrieval techniques including LAH[1] [28] and MuseHash [29]. The number of epochs (10, 50, 100, 150) for each volumetric method (MeshNet[2] [15], MeshCNN [3] [16], DGCNN [4] [13], PointNet[5] [14], ResNet50[6] [25], GoogleNet[7] [26], CMCL[8] [20], CMIC[9] [21]) during MAP computation is also varied. We adhere to recommended training and testing sizes specified by the authors [31], [30] for each dataset. Employing a 5-fold cross-validation methodology ensures robustness across all experiments. In subsequent tables, the '*' symbol indicates that MuseHash exhibits statistical significance compared to other methods, as confirmed by a t-test.

[1]https://github.com/IDSM-AI/LAH
[2]https://github.com/iMoonLab/MeshNet
[3]https://github.com/ranahanocka/MeshCNN
[4]https://github.com/antao97/dgcnn.pytorch
[5]https://github.com/yanx27/Pointnet_Pointnet2_pytorch
[6]https://github.com/JayPatwardhan/ResNet-PyTorch
[7]https://github.com/Lornatang/GoogLeNet-PyTorch
[8]https://github.com/LongLong-Jing/Cross-Modal-Center-Loss
[9]https://github.com/IGLICT/IBSR_jittor

B. Experiment I: Analysis of Individual Modalities

In this section, we revisit experiments outlined in prior research [22], where we utilised MeshNet and MeshCNN for the mesh modality, and combined CMCL, LAH, and MuseHash with MeshNet for mesh based queries. Moreover, we extend upon this previous work by introducing additional methods and employing alternative models for point cloud and image modalities. Additionally, we present the results obtained using individual modalities and MAP metric, focusing on meshes, point clouds and images separately.

Table I contains the experiments based only on the mesh modality. MeshNet outpaces MeshCNN in mesh based queries. The methods incorporated with MeshNet consistently outperformed those with MeshCNN around by 2% in each dataset. This is likely due to its superior ability to capture intricate geometric features and spatial relationships within mesh data. Moreover, MuseHash outperforms all methods in these experiments.

In Table II, the experiments focusing solely on the point cloud modality. When only utilising point cloud data, DGCNN tends to enhance methods more than PointNet about 0.5% for BNv0 and 1% for MN40. similar to MeshNet in mesh analysis. MuseHash also outperforms all methods, especially with DGCNN.

Moving to Table III, evaluations of all models using the image modality are provided. Comparing ResNet50 and GoogleNet against CMCL, CMCI, MuseHash and LAH for 3D retrieval, GoogleNet consistently outperformed in all systems by approximately 5% for BNv0 and 9% for MN40, with MuseHash achieving the highest scores.

Overall, the experiments reaveal that utilising the image modality yields superior results compared to both mesh and point cloud modalities, with a 6% advantage over mesh and a 10% advantage over point cloud. This suggests the prominence of images for retrieval. Furthermore, MuseHash consistently excels across all modalities, showcasing its versatility. In the subsequent experiments, we consider the better representations; MeshNet, DGCNN and GoogleNet, for mesh, point cloud and image modalities, respectively.

C. Experiment II: Analysis of Combined Modalities

This section compares the performance of two modalities (MAP metric) against three modalities from the 3D retrieval domain (CMCL and CMCI), along with LAH and MuseHash from the image domain. Results are shown in Table IV and Table V. Table VI provides additional insights into precision, recall, and fscore metrics for CMCL, CMCI, and MuseHash using both two and three modalities. Overall, MuseHash consistently outperforms, especially with mesh and image modalities.

TABLE I
EXPERIMENT I-A: IMPACT OF MESH MODALITY ON THE MAP METRIC.

Dataset	No.	MeshCNN	MeshNet	CMCL [20]		CMIC [21]		Code	LAH [28]		MuseHash [29]	
	Epochs	[16]	[15]	MeshCNN	MeshNet	MeshCNN	MeshNet	Length	MeshCNN	MeshNet	MeshCNN	MeshNet
BNv0	10	0.6007*	0.6201*	0.6421*	0.6511*	0.6578*	0.6601*	16	0.7534*	0.7629*	0.7631*	**0.7723**
	50	0.6226*	0.6350*	0.6430*	0.6520*	0.6610*	0.6700*	32	0.7610*	0.7701*	0.7619*	**0.7791**
	100	0.6449*	0.6552*	0.6581*	0.6670*	0.6782*	0.6845*	64	0.7645*	0.7754*	0.7712*	**0.7834**
	150	0.6501*	0.6650*	0.6576*	0.6623*	0.6825*	0.6907*	128	0.7751*	0.7821*	0.7801*	**0.7883**
MN40	10	0.6726*	0.6801*	0.6902*	0.7097*	0.7202*	0.7351*	16	0.7790*	0.7811*	0.7891*	**0.8010**
	50	0.6900*	0.6954*	0.6934*	0.7099*	0.7315*	0.7466*	32	0.7840*	0.7889*	0.7903*	**0.8056**
	100	0.6711*	0.7091*	0.7048*	0.7103*	0.7435*	0.7506*	64	0.7945*	0.8001*	0.8032*	**0.8101**
	150	0.6502*	0.6654*	0.6594*	0.6595*	0.7505*	0.7610*	128	0.7910*	0.8058*	0.7998*	**0.8122**

TABLE II
EXPERIMENT I-B: IMPACT OF POINT CLOUD MODALITY ON THE MAP METRIC.

Dataset	No.	PointNet	DGCNN	CMCL [20]		CMIC [21]		Code	LAH [28]		MuseHash [29]	
	Epochs	[14]	[13]	PointNet	DGCNN	PointNet	DGCNN	Length	PointNet	DGCNN	PointNet	DGCNN
BNv0	10	0.5423*	0.5501*	0.6011*	0.6124*	0.6103*	0.6201*	16	0.7001*	0.7089*	0.7245*	**0.7313**
	50	0.5502*	0.5611*	0.6094*	0.6178*	0.6199*	0.6245*	32	0.7089*	0.7145*	0.7267*	**0.7377**
	100	0.5578*	0.5681*	0.6159*	0.6208*	0.6273*	0.6309*	64	0.7122*	0.7202*	0.7301*	**0.7399**
	150	0.5602*	0.5781*	0.6223*	0.6319*	0.6301*	0.6478*	128	0.7234*	0.7278*	0.7408*	**0.7489**
MN40	10	0.6345*	0.6401*	0.6511*	0.6590*	0.6634*	0.6710*	16	0.7025*	0.7110*	0.7314*	**0.7409**
	50	0.6402*	0.6556*	0.6601*	0.6712*	0.6745*	0.6588*	32	0.7098*	0.7183*	0.7380*	**0.7480**
	100	0.6551*	0.6590*	0.6645*	0.6701*	0.6789*	0.7001*	64	0.7134*	0.7209*	0.7401*	**0.7500**
	150	0.6423*	0.6501*	0.6590*	0.6688*	0.6790*	0.7045*	128	0.7184*	0.7256*	0.7446*	**0.7563**

TABLE III
EXPERIMENT I-C: IMPACT OF IMAGE MODALITY ON THE MAP METRIC.

Dataset	No.	ResNet50	GoogleNet	CMCL [20]		CMIC [21]		Code	LAH [28]		MuseHash [29]	
	Epochs	[25]	[26]	ResNet50	GoogleNet	ResNet50	GoogleNet	Length	ResNet50	GoogleNet	ResNet50	GoogleNet
BNv0	10	0.7078*	0.7101*	0.7301*	0.7423*	0.7489*	0.7578*	16	0.7710*	0.7801*	0.7910*	**0.8010**
	50	0.7123*	0.7289*	0.7420*	0.7589*	0.7545*	0.7645*	32	0.7801*	0.7909*	0.8001*	**0.8140**
	100	0.7212*	0.7321*	0.7505*	0.7611*	0.7601*	0.7725*	64	0.7910*	0.8009*	0.8123*	**0.8270**
	150	0.7301*	0.7401*	0.7680*	0.7666*	0.7798*	0.7800*	128	0.7959*	0.8123*	0.8240*	**0.8315**
MN40	10	0.7423*	0.7510*	0.7529*	0.7630*	0.7663*	0.7701*	16	0.7819*	0.7900*	0.8031*	**0.8151**
	50	0.7515*	0.7623*	0.7645*	0.7745*	0.7754*	0.7851*	32	0.7958*	0.8028*	0.8150*	**0.8245**
	100	0.7601*	0.7751*	0.7767*	0.7807*	0.7841*	0.7910*	64	0.8031*	0.8151*	0.8231*	**0.8339**
	150	0.7720*	0.7810*	0.7801*	0.7911*	0.7910*	0.8001*	128	0.8121*	0.8245*	0.8344*	**0.8461**

TABLE IV
EXPERIMENT II-A: IMPACT OF TWO MODALITIES. MESH (M), POINT CLOUD (PC), AND IMAGE (I) REPRESENTATIONS ON THE MAP METRIC.

Dataset	No.	CMCL [20]			CMIC [21]			Code	LAH [28]			MuseHash [29]		
	Epochs	M, PC	PC, I	M, I	M, PC	PC, I	M, I	Length	M, PC	PC, I	M, I	M, PC	PC, I	M, I
BNv0	10	0.6761*	0.6801*	0.7021*	0.6978*	0.7002*	0.7204*	16	0.7120*	0.7265*	0.7501*	0.7610*	0.7701*	**0.7910**
	50	0.6810*	0.6881*	0.7110*	0.7023*	0.7089*	0.7345*	32	0.7112*	0.7242*	0.7467*	0.7691*	0.7734*	**0.8012**
	100	0.6902*	0.6910*	0.7222*	0.7089*	0.7199*	0.7401*	64	0.7151*	0.7266*	0.7491*	0.7701*	0.7791*	**0.8110**
	150	0.6971*	0.7001*	0.7315*	0.7162*	0.7056*	0.7545*	128	0.7203*	0.7298*	0.7508*	0.7688*	0.7801*	**0.8291**
MN40	10	0.6910*	0.6710*	0.7011*	0.6834*	0.6885*	0.7123*	16	0.7311*	0.7381*	0.7395*	0.7882*	0.7712*	**0.8284**
	50	0.7039*	0.6912*	0.7110*	0.6904*	0.6941*	0.7223*	32	0.7362*	0.7398*	0.7405*	0.7910*	0.7821*	**0.8301**
	100	0.7128*	0.7010*	0.7222*	0.7011*	0.7071*	0.7389*	64	0.7384*	0.7401*	0.7421*	0.7900*	0.7791*	**0.8434**
	150	0.7231*	0.7122*	0.7515*	0.7045*	0.7101*	0.7448*	128	0.7392*	0.7412*	0.7433*	0.7854*	0.7840*	**0.8512**

Following the methodology used in the uni-modal case, we conducted experiments with CMCL, CMIC, and MuseHash. We then expanded to include two models each from GoogleNet, MeshNet, and DGCNN for combined two-modalities queries (Table IV). Combining both mesh and image modalities yielded the most promising results, followed by using point cloud and image modalities (Table IV). However, relying solely on mesh or point cloud modalities resulted in lower outcomes, likely due to the absence of finer details in point clouds compared to meshes, which limits their representation of complex object structures. Moreover, using only one modality led to a decrease of approximately 6%

TABLE V
EXPERIMENT II-B: IMPACT OF THREE MODALITIES. MESH (M), POINT CLOUD (PC), AND IMAGE (I) REPRESENTATIONS ON THE MAP METRIC.

Dataset	No. Epochs	CMCL [20] M, PC, I	CMIC [21] M, PC, I	Code Length	MuseHash [29] M, PC, I
BNv0	10	0.6611*	0.7111*	16	**0.7870**
	50	0.6720*	0.7220*	32	**0.7920**
	100	0.6870*	0.7364*	64	**0.8001**
	150	0.6923*	0.7400*	128	**0.8032**
MN40	10	0.7197*	0.7069*	16	**0.8101**
	50	0.7289*	0.7131*	32	**0.8236**
	100	0.7343*	0.7264*	64	**0.8323**
	150	0.7499*	0.7345*	128	**0.8401**

compared to exclusively using two modalities. This underscores the importance of leveraging complementary modalities to enhance the efficacy of 3D retrieval.

Next, we explore the utilisation of all networks for combined three-modalities queries (Table V). When incorporating all modalities, we observe a slight decrease of approximately 2% compared to using only two modalities across all methods and datasets (Table IV). This indicates that while the inclusion of additional modalities provides more information, it also introduces some complexity that may slightly impact overall performance. However, despite this slight decrease, leveraging all available modalities remains valuable as it allows for a more comprehensive representation of the data and potentially improves retrieval results in certain scenarios.

Regarding the experimental analysis presented in Table VI, we gain valuable insights into the performance of retrieval methods, particularly in ranking and retrieving relevant items. This examination covers multimodal approaches from the 3D domain, such as CMCL and CMIC, evaluated across various numbers of epochs. Additionally, MuseHash from the image domain undergoes thorough evaluation across different hash code lengths. Furthermore, the study investigates the use of both two modalities and three modalities for each method. In uni-modal scenarios, where individual modalities are considered, MeshNet and DGCNN emerge as the optimal models for the mesh and point cloud modalities, respectively.

While CMCL may perform well at times, MuseHash is generally more efficient , with around a 10% higher Fscore. MuseHash's ability to combine multiple modalities into a single hash code improves retrieval accuracy and is particularly useful for large datasets and resource-limited environments where quick and accurate searches are essential.

In summary, MuseHash excels in multimodal retrieval, particularly with mesh and image modalities. While combining mesh and image modalities yields good results, relying solely on mesh or point cloud leads to lower outcomes. Despite some complexity and a slight performance decrease, using all modalities remains valuable for comprehensive data representation.

D. Conclusive Insights

In conclusion, this study offers significant insights into the fusion of different types of information for 3D retrieval tasks. Figure 2 showcases the optimal outcomes achieved by MuseHash on BuildingNet_v0 (highlighted in green) and ModelNet40 (highlighted in blue), with performance metrics the integer part of MAP multiplied by 100%. A Venn diagram is employed to depict the varying MAP values resulting from different combinations of modalities. Through experimentation with various descriptor combinations, it becomes apparent that the use of mesh and image data is advantageous (~ 82% in BuildingNet_v0 and ~ 85% in ModelNet40). In particular, the utilisation of MeshNet from mesh based models and GoogleNet from image based models significantly enhances results, benefiting not only MuseHash but also all other methods. Moreover, the superior performance of MuseHash, evidenced by consistently higher Fscores and MAP values compared to other methods, underscores its effectiveness. These findings hold promise for researchers and practitioners seeking to refine the search process for 3D objects, enhancing efficiency and accuracy in various applications.

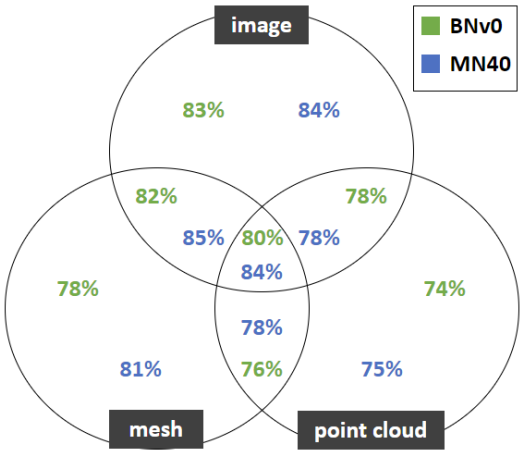

Fig. 2. Overview of MuseHash's performance across the modalities.

V. CONCLUSION

Recent research emphasizes the significance of managing diverse 3D representations, and our study addresses this by conducting point cloud and image experiments in the uni-modal scenario. Through experimenting with various descriptor combinations, we find that utilising both mesh and image data is advantageous. Specifically, leveraging MeshNet from mesh based models and GoogleNet from image based models notably improves results, benefiting not only MuseHash but all other methods as well. Furthermore, the consistently higher fscores and MAP values of MuseHash compared to other methods underscore its effectiveness. These findings offer promise for researchers and practitioners seeking to optimize the search process for 3D objects, enhancing efficiency and accuracy across various applications.

ACKNOWLEDGMENT

This work was supported by the Horizon research and innovation programme under grant agreement HE-101070250 XRECO.

TABLE VI

EXPERIMENT II-C: COMPARING METHODS FOR PRECISION AT K, RECALL AT K AND F-SCORE AT K (K = 10, 25, 50) WITH VARYING EPOCHS OR CODE LENGTHS ON MN40 DATASET. MESH (M), POINT CLOUD (PC), AND IMAGE (I) REPRESENTATIONS

Method	Modality	Variable	Precision@k			Recall@k			Fscore@k		
		Epochs	**10**	**25**	**50**	**10**	**25**	**50**	**10**	**25**	**50**
CMCL [20]	*M*	10	0.7623	0.7631	0.7670	0.9101	0.9121	0.9132	0.8297	0.8310	0.8337
	PC	50	0.7630	0.7638	0.7671	0.9120	0.9124	0.9140	0.8309	0.8315	0.8341
		100	0.7522	0.7520	0.7412	0.9134	0.9155	0.9169	0.8250	0.8257	0.8197
		150	0.7501	0.7482	0.7389	0.9201	0.9231	0.9251	0.8264	0.8265	0.8216
	PC	10	0.7810	0.7823	0.7830	0.9010	0.9041	0.9045	0.8367	0.8388	0.8394
	I	50	0.7831	0.7828	0.7835	0.9017	0.9049	0.9050	0.8382	0.8394	0.8399
		100	0.7840	0.7832	0.7840	0.9020	0.9050	0.9055	0.8389	0.8397	0.8404
		150	0.7851	0.7848	0.7850	0.9021	0.9025	0.9030	0.8395	0.8395	0.8399
	M	10	0.7910	0.7954	0.7970	0.9011	0.9045	0.9051	0.8425	0.8464	0.8476
	I	50	0.7921	0.7961	0.7982	0.9051	0.9062	0.9076	0.8448	0.8476	0.8494
		100	0.7933	0.7972	0.7990	0.9060	0.9079	0.9081	0.8459	0.8490	0.8501
		150	0.7941	0.7980	0.7921	0.9071	0.9077	0.9091	0.8468	0.8493	0.8466
	M	10	0.8290	0.7679	0.7142	**0.9985**	0.9943	0.9968	0.9011	0.8666	0.8321
	PC	50	0.8291	0.7687	0.7147	0.9883	0.9943	**0.9968**	0.9018	0.8671	0.8325
	I	100	**0.8298**	0.7687	0.7149	0.9884	**0.9944**	**0.9968**	**0.9019**	0.8671	0.8326
		150	0.8283	0.7677	0.7142	0.9865	**0.9944**	**0.9968**	0.9013	0.8665	0.8322
CMIC [21]	*M*	10	0.5712	0.5721	0.5745	0.9678	0.9699	0.9712	0.7184	0.7197	0.7219
	PC	50	0.5720	0.5739	0.5791	0.9699	0.9738	0.9756	0.7196	0.7222	0.7268
		100	0.5731	0.5744	0.5799	0.9701	0.9712	0.9788	0.7205	0.7219	0.7283
		150	0.5748	0.5771	0.5801	0.9723	0.9740	0.9792	0.7225	0.7248	0.7286
	PC	10	0.5882	0.5888	0.5890	0.9710	0.9732	0.9756	0.7326	0.7337	0.7345
	I	50	0.5890	0.5912	0.5954	0.9723	0.9744	0.9789	0.7336	0.7359	0.7404
		100	0.5901	0.5929	0.5998	0.9731	0.9751	0.9791	0.7354	0.7374	0.7439
		150	0.5923	0.5939	0.6012	0.9744	0.9770	0.9810	0.7368	0.7387	0.7455
	M	10	0.6021	0.6034	0.6042	0.9601	0.9682	0.9699	0.7401	0.7435	0.7446
	I	50	0.6034	0.6041	0.6052	0.9610	0.9690	0.9700	0.7413	0.7442	0.7454
		100	0.6045	0.6078	0.6101	0.9651	0.9660	0.9711	0.7437	0.7461	0.7494
		150	0.6052	0.6091	0.6132	0.9689	0.9691	0.9719	0.7450	0.7480	0.7520
	M	10	0.5910	0.5920	0.5931	0.9601	0.9623	0.9630	0.7316	0.7330	0.7341
	PC	50	0.5931	0.5942	0.5960	0.9620	0.9632	0.9643	0.7338	0.7350	0.7367
	I	100	0.5940	0.5949	0.5971	0.9637	0.9642	0.9652	0.7350	0.7358	0.7378
		150	0.5952	0.5958	0.5980	0.9642	0.9651	0.9660	0.7360	0.7368	0.7387
		Code Length	**10**	**25**	**50**	**10**	**25**	**50**	**10**	**25**	**50**
MuseHash [29]	*M*	16	0.6321	0.6389	0.6401	0.9456	0.9489	0.9490	0.7577	0.7636	0.7645
	PC	32	0.6345	0.6391	0.6435	0.9591	0.9501	0.9537	0.7637	0.7642	0.7685
		64	0.6401	0.6420	0.6456	0.9599	0.9771	0.9601	0.7680	0.7749	0.7721
		128	0.6411	0.6429	0.6481	0.9600	0.9611	0.9651	0.7688	0.7704	0.7755
	PC	16	0.6421	0.6589	0.6701	0.9521	0.9567	0.9600	0.7670	0.7804	0.7893
	I	32	0.6490	0.6601	0.6788	0.9589	0.9551	0.9663	0.7741	0.7807	0.7974
		64	0.6510	0.6678	0.6791	0.9601	0.9634	0.9691	0.7759	0.7888	0.7986
		128	0.6601	0.6702	0.6821	0.9678	0.9681	0.9700	0.7849	0.7921	0.8010
	M	16	0.6571	0.6810	0.7020	0.9612	0.9723	0.9865	0.7806	0.8010	0.8203
	I	32	0.6910	0.7001	0.7112	0.9546	0.9612	0.9667	0.8017	0.8101	0.8195
		64	0.7662	0.7405	0.7156	0.9712	0.9781	0.9801	0.8566	0.8429	0.8272
		128	0.8010	**0.8588**	**0.8423**	0.9865	0.9902	0.9923	0.8841	**0.9198**	**0.9112**
	M	16	0.6480	0.6501	0.6589	0.9523	0.9678	0.9621	0.7712	0.7778	0.7821
	PC	32	0.6510	0.6678	0.6781	0.9678	0.9698	0.9512	0.7784	0.7910	0.7918
	I	64	0.6782	0.6789	0.6834	0.9701	0.9700	0.9634	0.7983	0.7988	0.7996
		128	0.7012	0.6910	0.6901	0.9701	0.9623	0.9603	0.8140	0.8044	0.8038

References

[1] E. Mohr, T. Thum, and C. Bär, "Accelerating Cardiovascular Research: Recent Advances in Translational 2D and 3D Heart Models," *European Journal of Heart Failure*, vol. 24, no. 10, pp. 1778–1791, 2022.

[2] M. Javaid, A. Haleem, R. P. Singh, and R. Suman, "Industrial Perspectives of 3D Scanning: Features, Roles and it's Analytical Applications," *Sensors International*, vol. 2, 2021.

[3] M. M. Dummer, K. L. Johnson, S. Rothwell, K. Tatah, and M. K. Hibbs-Brenner, "The Role of VCSELs in 3D Sensing and LiDAR," in *OPTO*, 2021.

[4] M. J. Willemink, W. A. Koszek, C. Hardell, J. Wu, D. Fleischmann, H. Harvey, L. R. Folio, R. M. Summers, D. L. Rubin, and M. P. Lungren, "Preparing Medical Imaging Data for Machine Learning," *Radiology*, vol. 295, no. 1, pp. 4–15, 2020.

[5] M. L. Brutto and P. Meli, "Computer Vision Tools for 3D Modelling in Archaeology," in *hde*, 2012.

[6] Y.-S. Han, J. Lee, J. Lee, W. Lee, and K. Lee, "3D CAD Data Extraction and Conversion for Application of Augmented/Virtual Reality to the Construction of Ships and Offshore Structures," *International Journal of Computer Integrated Manufacturing*, vol. 32, no. 7, pp. 658–668, 2019.

[7] Q. Ha, L. Yen, and C. Balaguer, "Robotic Autonomous Systems for Earthmoving in Military Applications," *Automation in Construction*, vol. 107, 2019.

[8] P. Song, Y. Qi, and D. Cai, "Research and Application of Autodesk Fusion360 in Industrial Design," in *IOP Conference Series: Materials Science and Engineering*, vol. 359, no. 1. IOP Publishing, 2018, p. 012037.

[9] H. T. G. Le, "3D Modeling Assets And Props With Maya: General 3D Modeling Pipeline," 2022.

[10] C. Morse, "Gaming Engines: Unity, Unreal, and Interactive 3D Spaces," 2021.

[11] H. L. Jacobs, "SketchUp and Sketchfab: Tools for Teaching with 3D," *Journal of the Society of Architectural Historians*, vol. 81, no. 2, pp. 256–259, 2022.

[12] E. A. Juanda and F. Khairullah, "Tinkercad Application Software to Optimize Teaching and Learning Process in Electronics and Microprocessors Subject," in *6th UPI International Conference on TVET 2020 (TVET 2020)*. Atlantis Press, 2021, pp. 124–128.

[13] Y. Wang, Y. Sun, Z. Liu, S. E. Sarma, M. M. Bronstein, and J. M. Solomon, "Dynamic Graph CNN for Learning on Point Clouds," *ACM Transactions on Graphics (tog)*, vol. 38, no. 5, pp. 1–12, 2019.

[14] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep Learning on Point Sets for 3D Classification and Segmentation," in *CVPR*, Honolulu, HI, USA, 2017.

[15] Y. Feng, Y. Feng, H. You, X. Zhao, and Y. Gao, "MeshNet: Mesh Neural Network for 3D Shape Representation," in *AAAI*, Honolulu, Hawaii, 2019.

[16] R. Hanocka, A. Hertz, N. Fish, R. Giryes, S. Fleishman, and D. Cohen-Or, "MeshCNN: A Network with an Edge," *ACM Transactions on Graphics (ToG)*, vol. 38, no. 4, pp. 1–12, 2019.

[17] D. Lin, Y. Li, Y. Cheng, S. Prasad, T. L. Nwe, S. Dong, and A. Guo, "Multi-view 3D Object Retrieval Leveraging the Aggregation of View and Instance Attentive Features," *Knowledge-Based Systems*, vol. 247, p. 108754, 2022.

[18] J.-C. Su, M. Gadelha, R. Wang, and S. Maji, "A Deeper Look at 3D Shape Classifiers," in *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 2018.

[19] A. S. Gezawa, Y. Zhang, Q. Wang, and L. Yunqi, "A Review on Deep Learning Approaches for 3D Data Representations in Retrieval and Classifications," *IEEE Access*, vol. 8, pp. 57 566–57 593, 2020.

[20] L. Jing, E. Vahdani, J. Tan, and Y. Tian, "Cross-Modal Center Loss for 3D Cross-Modal Retrieval," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3142–3151.

[21] M.-X. Lin, J. Yang, H. Wang, Y.-K. Lai, R. Jia, B. Zhao, and L. Gao, "Single Image 3D Shape Retrieval via Cross-Modal Instance and Category Contrastive Learning," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 11 405–11 415.

[22] M. Pegia, B. Þ. Jónsson, A. Moumtzidou, S. Diplaris, I. Gialampoukidis, S. Vrochidis, and I. Kompatsiaris, "Multimodal 3D Object Retrieval," in *International Conference on Multimedia Modeling*. Springer, 2024, pp. 188–201.

[23] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[25] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.

[27] Z. Gao, Y. Li, and S. Wan, "Exploring Deep Learning for View-based 3D Model Retrieval," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 16, no. 1, pp. 1–21, 2020.

[28] Y. Xie, Y. Liu, Y. Wang, L. Gao, P. Wang, and K. Zhou, "Label-Attended Hashing for Multi-Label Image Retrieval." in *IJCAI*, 2020, pp. 955–962.

[29] M. Pegia, B. Þ. Jónsson, A. Moumtzidou, I. Gialampoukidis, S. Vrochidis, and I. Kompatsiaris, "MuseHash: Supervised Bayesian Hashing for Multimodal Image Representation," in *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*, 2023, pp. 434–442.

[30] P. Selvaraju, M. Nabail, M. Loizou, M. Maslioukova, M. Averkiou, A. Andreou, S. Chaudhuri, and E. Kalogerakis, "BuildingNet: Learning to Label 3D Buildings," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 397–10 407.

[31] Z. Wu, S. Song, A. Khosla, F. Yu, L. Zhang, X. Tang, and J. Xiao, "3D Shapenets: A Deep Representation for Volumetric Shapes," in *CVPR*, 2015, pp. 1912–1920.