Multimodality in Media Retrieval

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

ABSTRACT

The quest for retrieving relevant media for a given query is well-studied and has various applications. Modern publicly available media collections provide diverse modalities of the same objects, which can enhance search. Our research delves into enhancing media retrieval by effectively representing and querying multimodal data. In the retrieval methods' ranking procedure, we examine efficiency through techniques like approximate nearest neighbor (ANN) indexing and high-performance computing (HPC). Our method, MuseHash, is proposed for single media object retrieval and is applied to images and 3D objects, outperforming existing methods on diverse datasets. Moreover, it significantly reduces execution times with ANN and HPC. Future plans include considering multimodality in the video retrieval domain.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Information retrieval query processing**; • **Human-centered computing** → **Visualization systems and tools**; • **Computing methodologies** → **Supervised learning**.

KEYWORDS

Media Retrieval, Multimodality, Optimization

ACM Reference Format:
Maria Eirini Pegia. 2024. Multimodality in Media Retrieval. In *Proceedings of the 2024 International Conference on Multimedia Retrieval (ICMR '24), June 10–14, 2024, Phuket, Thailand.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3652583.3657583

1 INTRODUCTION

In today's digital era, the Internet grants us easy access to a plethora of media, ranging from simple images and text to more intricate structures like videos and 3D graphics. For instance, when viewing a scene from our favorite series, we perceive it as a dynamic structure comprising a sequence of frames, each accompanied by captions, timestamps, locations, and audio elements. Moreover, within the video, characters exist within a 3D space, adding layers of complexity. Despite its complexity, each video component corresponds to a specific moment.

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
ICMR '24, June 10–14, 2024, Phuket, Thailand
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0619-6/24/06.
https://doi.org/10.1145/3652583.3657583

These diverse forms of media find applications across various domains, including urban development [8], gaming, healthcare [37], historical analysis [25], archaeology [6] and computer-aided design (CAD) [12]. However, the challenge lies in efficiently representing and amalgamating this heterogeneous data. Therefore, our primary aim is to devise methods for effectively representing and integrating [1, 2, 10, 13, 40] the information within these media collections.

To achieve this aim, we categorise different information of media into distinct abstract views, known as modalities. Our research focuses on four primary modalities: visual, which encompasses high-resolution images (VHR), low resolution images from drones [18, 34] and individual frames extracted from videos; text, including keyframes of images and complex captions generated by models like CLIP [32]; temporal, referring to timestamps or timeframes within videos, capturing the temporal aspect of the data; and spatial, involving geographical coordinates and intricate structures such as point clouds representing objects in 3D space [26]. While other modalities exist, such as in medical applications, our current research primarily emphasizes these four. This classification enables us to develop a structured approach to represent and analyse diverse media collections effectively.

By comprehending and effectively incorporating these modalities, we aim to enhance the management and analysis of diverse media collections. As we navigate these challenges, our research direction encompasses the following objectives:

Objective 1 Developing multimodal retrieval approaches for static moments, spanning from simple image collections to complex 3D datasets (Section 3.1).

Objective 2 Exploring multimodal methods in large-scale realistic scenarios via indexing and query processing optimization (Section 3.2).

Objective 3 Exploring multimodal methods in video retrieval. This will be outlined as future steps in Section 4.

To address the first objective, we created MuseHash [31], a supervised Bayesian framework for unimodal image retrieval. We then advanced MuseHash [29] to facilitate multimodal retrieval, a development lauded for its adaptability across datasets and recognised excellence in 3D object retrieval [28].

For the second objective, we have connected ANN methods, MuseHash, and High-Performance Computing (HPC) infrastructures [30]. Our observations indicate that certain ANN methods outperform brute-force ranking approaches. GPUs show potential for longer hashes due to their parallel processing capabilities. Our research underscores the superiority of query parallelism over data parallelism in retrieval strategies.

The remainder of this paper is organized as follows: Section 2 delves into previous works in the field. Section 3 provides an

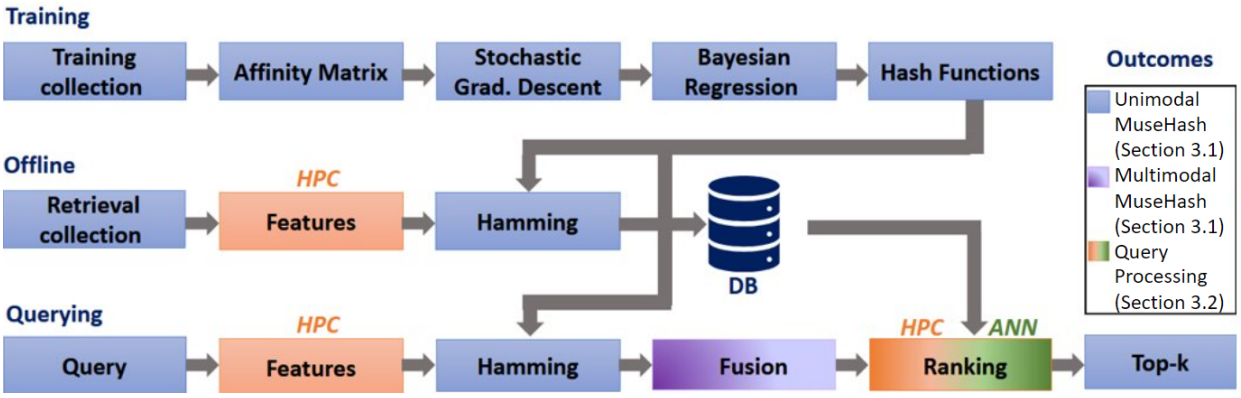

Figure 1: Overall outcomes.

overview of our methodologies and outcomes on multimodality in image retrieval (Section 3.1) as well as our query processing evaluation techniques and its outcomes (Section 3.2). Our research plans for the next year are outlined in Section 4. Finally, Section 5 concludes the paper.

2 BACKGROUND

In this section, we overview current state-of-the-art methods in our research field, including multimodal retrieval techniques (Section 2.1) and query processing evaluation methods (Section 2.2).

2.1 Cutting-edge Multimodal Retrieval Techniques

In our investigation of multimodality in image retrieval, we categorise it into unimodal and multimodal scenarios based on the number of modalities involved. We prioritise supervised techniques, especially supervised hashing methods, known for their superior retrieval accuracy compared to unsupervised methods, along with their memory efficiency and speed in the retrieval process.

In unimodal image retrieval, a single modality, typically one specific type of data, is utilized. Our study explores various modalities such as text, image, datetime, location, mesh, and point-cloud, investigating diverse retrieval scenarios and applications. Notably, modalities like datetime and location have not received as much research attention as image and text modalities.

Supervised methods in this context often leverage deep learning networks like Convolutional Neural Networks (CNNs) for feature learning and hash function development. Deep Cauchy Hashing (DCH) [7] optimizes hash codes using Cauchy cross-entropy loss and quantization loss within a deep learning framework. Semantic Preserving Hashing (SePH) [21] minimizes Kullback-Leibler divergence to approximate semantic affinities while unifying hash codes for various views.

We surveyed various supervised hashing methods for multimodal image retrieval, which encompass strategies such as similarity-based, adversarial-based, deep neural networks, and discrete-based methods. FCMH [36] optimizes binary codes and DOCH [41] generates high-quality hash codes. LAH [39] focuses on image representations and label co-occurrence embeddings with Cauchy distribution-based hash functions. SSAH [19] incorporates a self-supervised semantic network and adversarial learning. GSPH [24] learns hash codes and functions for two modalities. MTHF [22] transfers knowledge from single-modal to cross-modal domains, while KDLFH [20] directly learns binary hash codes.

While some methods are tailored for cross-modal scenarios, exceptions like LAH [39] support multimodal queries. In the 3D retrieval domain, CMCL [15] integrates multiple 3D modalities but can be computationally intensive and dataset-sensitive.

2.2 Cutting-edge Optimization Methods

We selected several cutting-edge ANN methods based on their unique and complementary features, following the approach outlined by Aumüller et al. [3]. These include tree-based structures, graph-based structures, pruning techniques, brute-force approaches, and baseline methods.

3 CURRENT RESEARCH RESULTS

Figure 1 visually summarizes our research outcomes, each linked to its respective section for more details. To simplify, the light blue boxes represent the unimodal MuseHash method [31], which is the foundation of our work. The light purple blocks signify the extended MuseHash method [29], to handle multimodal data. The light orange and green blocks represent the investigations and evaluations in query processing, introducing two new components to our research.

3.1 Multimodality in Multimodal Retrieval

3.1.1 Methodology. In our research, we created a novel supervised hashing method called MuseHash [31]. MuseHash leverages Bayesian principles for hash function learning and adapts to the data's statistical properties, enhancing overall hashing and retrieval system performance. The method comprises three main phases: training, offline, and querying, each illustrated with light blue boxes in Figure 1.

During training, hash functions are generated from the training collection via Bayesian ridge regression, mapping feature vectors from the visual modality to the Hamming space. Affinity matrices are created using both ground truth labels and cosine similarity,

Table 1: MAP and accuracy results for ModelNet40 and BuildingNet_v0 with different code lengths or number of epochs and query modalities.

Dataset	Query	No. Epochs	CMCL[15] mAP	CMCL[15] Accuracy	Hash Length	MuseHash [29] mAP	MuseHash [29] Accuracy
ModelNet40	**Mesh**	10	0.7097*	0.7916*	16	**0.8010**	**0.9431**
		50	0.7099*	0.8001*	32	**0.8056**	**0.9488**
		100	0.7103*	**0.9791**	64	**0.8101**	0.9500
		150	0.6695*	**0.9895**	128	**0.8122**	0.9510
	Visual Mesh	10	0.6911*	0.9012*	16	**0.8184**	**0.9501**
		50	0.7010*	0.9045*	32	**0.8201**	**0.9578**
		100	0.7122*	0.9091*	64	**0.8234**	**0.9601**
		150	0.7415*	0.9129*	128	**0.8212**	**0.9525**
	Visual Mesh Point Cloud	10	0.7097*	0.7916*	16	**0.8051**	**0.9611**
		50	0.7099*	0.8001*	32	**0.7976**	**0.9601**
		100	0.7103*	**0.9791**	64	**0.7923**	0.9583
		150	0.6695*	**0.9895**	128	**0.7911**	0.9550

Table 2: Multimodal query results for different datasets and different code lengths and all modalities.

Dataset	Method	16bit	32bit	64bit	128bit
AU-AIR	LAH	0.9054*	0.8723	0.8700*	0.8821*
	MuseHash	**0.9726**	**0.8790**	**0.8861**	**0.9019**
MarDCT	LAH	0.7511*	0.7601*	0.7710*	0.7765*
	MuseHash	**0.7803**	**0.7811**	**0.7851**	**0.7899**
SeaDronesSee	LAH	0.8422*	0.8450*	0.8501*	0.8607*
	MuseHash	**0.8626**	**0.8690**	**0.8741**	**0.8819**
Mirflickr25K	LAH	0.8233*	0.8309*	0.8440*	0.8401*
	MuseHash	**0.8503**	**0.8541**	**0.8551**	**0.8599**
NUS-WIDE	LAH	0.8513*	0.8609*	0.8743*	0.8823*
	MuseHash	**0.9303**	**0.9341**	**0.9381**	**0.9409**

from which semantic probabilities are derived through normalization. In the offline phase, features are extracted from the retrieval set for the visual modality. Using the learned hash functions, hash codes are computed and stored in a database, ensuring efficient storage and retrieval of multimedia data. Finally, during the querying phase, the learned hash functions are applied to a given query, and the database is queried using Hamming distances to retrieve the top-k relevant results.

To address flexible multimodal approaches, we extended MuseHash [31] to support any number of modalities [29]. This enhancement enables efficient fusion of different modalities for the same object, computing hash codes and retrieving relevant items.

3.1.2 Evaluation. The researchers in the retrieval domain are more familiar with very high resolution data, rather than data from underwater and aerial footage. It is a challenge to handle all this different collections in such a way that the retrieval is efficient.

Hence, MuseHash [29] is specifically is designed specifically for multimodal queries across diverse collections. In our study, we compare MuseHash with LAH across five datasets (Table 2), varying hash code lengths and utilizing all available modalities. In specific table, both the query and each element within the collection leverage all available modalities, integrating them into a unified hash code. While LAH can only fuse two modalities, MuseHash accommodates more than two modalities.

AU-AIR [5] and MarDCT [4] are UAV datasets that contains image with temporal, and image with geotemporal information, respectively. SeaDronesSee [35] is a underwater dataset, which include images with geotemporal information. MIRFlickr25K [14] and NUS-WIDE [9] are benchmark high resolution datasets commonly utilized in the literature.

MuseHash consistently outperforms LAH in all cases with statistical significance (the symbol "*" denotes statistical significance after t-test). For enhanced evaluation robustness, we utilized a 5-fold cross-validation methodology across all experiments. Particularly, MuseHash performs exceptionally well when using all modalities in the MarDCT, MIRFlickr25K, and NUS-WIDE datasets, benefiting from the high-quality information present in these collections.

In summary, MuseHash emerges as a superior performer, consistently outpacing seven state-of-the-art methods across diverse image collections in both multimodal and unimodal scenarios. It achieves this by leveraging a combination of various visual descriptors such as VGG16 and ResNet50, and textual descriptors like Bag-of-Words (BoW) or BERT. Moreover, MuseHash offers flexibility with hash code lengths ranging from 16-bit to 128-bit, catering to different requirements and scenarios. This comprehensive approach enables MuseHash to exhibit greater robustness compared to existing methods, making it a compelling choice for multimodal query tasks across diverse datasets.

Expanding into 3D collections, we have extended the application of MuseHash from image retrieval to 3D object retrieval [28]. Specifically, we have adapted the multimodal MuseHash technique for volumetric data queries. In this context, the multimodal approach integrates various types of data representations associated with 3D objects, such as meshes, point clouds. By doing so, MuseHash extends its utility beyond traditional 2D image datasets, allowing

for more comprehensive and effective retrieval of 3D objects based on multimodal queries.

During evaluation, MuseHash consistently outperformed other methods in both unimodal and multimodal scenarios across different hash lengths and epochs for the ModelNet40 dataset (Table 1). ModelNet40 [38] and BuildingNet_v0 [33] are two publicly available benchmark datasets, containing image, mesh, and point cloud representations of 3D objects. The former is a dataset of 3D CAD models dedicated to object categories (e.g., car, airplane), while the latter includes different building types (e.g., church, palace) associated with their textures (e.g., color, material).

While the CMCL approach excelled in accuracy with more epochs, its mAP performance fell short (similarly, the symbol "*" denotes statistical significance after t-test). Likewise, we conducted 5-fold-cross validation in the evaluation. Overall, MuseHash demonstrated competitive performance. The multimodal variant showed significant performance improvements with longer code lengths (16 to 32), especially for larger lengths (64 and 128). However, extending the code length beyond this range did not yield substantial gains.

In conclusion, our study applied advanced image retrieval methods to 3D object retrieval, adapting MuseHash for volumetric data queries. MuseHash's exploitation of inter-modality relationships consistently outperformed three state-of-the-art methods across two benchmark image collections.

3.2 Evaluation in Query Processing

3.2.1 Methodology. This section discusses the relationship between MuseHash, Approximate Nearest Neighbor (ANN) methods, and High-Performance Computing (HPC) Infrastructure [30]. Figure 1 illustrates new components integrated into the MuseHash architecture, depicted in light green and light orange. The light green block represents the integration of ANN methods into the MuseHash ranking process, while the light orange block signifies optimization techniques applied to feature and ranking processes.

We prioritise hardware resource optimization through parallel computing using multithreading and GPUs, particularly leveraging NVIDIA CUDA for efficient feature extraction in both offline and querying phases. Additionally, we explore multi-GPU setups to significantly enhance performance. To evaluate our methods' performance, we utilize two approaches: data parallelism and query parallelism.

Data parallelism Splitting the data into segments enables distribution among multiple processes for searching, allowing us to assess system scalability when numerous processes collaborate to process the data.

Query parallelism Maintaining a pool of processes ready to handle incoming queries allows for efficient allocation of queries, aiding in the assessment of how effectively the system manages concurrent queries.

3.2.2 Evaluation. When handling multimodal media, the speed and efficiency are crucial. Our objective [30] is to identify effective techniques for data acquisition and analysis across both CPU and GPU platforms. The CPU experiments encompass datasets such as AU-AIR and LSC'23 [11] datasets, while the GPU experiments focusing solely on AU-AIR.

The LSC'23 dataset was generated by an active lifelogger over the course of 18 months and captured by a wearable camera. Every image within this dataset is associated with pertinent captions, temporal data, spatial information, or a combination of these elements. Below, we summarize our main findings.

Superior Throughput and Quality The graph-based Hnswlib method excelled in both throughput and result quality

Multi-Core Synergy Combining specific data organization methods with utilizing multiple cores of a computer concurrently results in accelerated processing.

Comparison of GPU and CPU CPUs beat GPUs in some tasks, needing smart strategies for specialised chip potential. Complex data slows processing, crucial to tackle for efficiency.

4 FUTURE WORK

Our plans for the remainder of the thesis involves developing an efficient video data querying framework. Initially, we aim to devise a method tailored to benchmark datasets and benchmark it against state-of-the-art (SOTA) methods. Furthermore, we plan to evaluate video retrieval methods using more realistic workloads and datasets like the Known-Item Search (KIS) tasks from Video Browser Showdown (VBS)[23] and Lifelog Search Challenge (LSC) [11] competitions. Additionally, inspired by recent research, we seek to explore various modalities for the VBS task, investigating which combinations—such as action-based, visual-based, or CLIP-based—yield the most benefit for our system[16, 17]. If initial approaches do not yield expected results, we consider examining SOTA reinforcement learning methods for the KIS task as an alternative direction. Subsequently, we intend to integrate our proposed techniques into our search engine system, VERGE [27], which participates in the VBS competition. This will allow us to observe first-hand how our framework performs in practical search scenarios.

5 CONCLUSION

In our multimodal image retrieval research, MuseHash, consistently outperforms state-of-the-art methods in uni-modal and multimodal scenarios. Extending MuseHash to 3D object retrieval demonstrates its versatility with volumetric data. We highlight the efficiency of ANN methods over brute-force approaches. Scalability experiments reveal GPUs' potential for longer hashes. Query parallelism surpasses data parallelism in retrieval strategies, enhancing performance and adaptability. Finally, we conclude with some future research direclintions in video retrieval field.

ACKNOWLEDGMENT

This work was supported by the Horizon research and innovation programme under grant agreement HE-101070250 XRECO and the European Security Funds under the topic ISFP-2021-AG-TCO from the EU Home Affairs under grant agreement 101080090 ALLIES.

REFERENCES

[1] Charu C Aggarwal. 2018. Information retrieval and search engines. *Machine Learning for Text* (2018), 259–304.

[2] Giuseppe Amato, Fabio Carrara, Fabrizio Falchi, and Claudio Gennaro. 2017. Efficient indexing of regional maximum activations of convolutions using full-text search engines. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval*. 420–423.

[3] Martin Aumüller, Erik Bernhardsson, and Alexander Faithfull. 2017. ANN-Benchmarks: A Benchmarking Tool for Approximate Nearest Neighbor Algorithms. In *International Conference on Similarity Search and Applications*. Springer, 34–49.

[4] Domenico D Bloisi, Luca Iocchi, Andrea Pennisi, and Luigi Tombolini. 2015. ARGOS-Venice Boat Classification. In *2015 12th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*. IEEE, 1–6.

[5] Ilker Bozcan and Erdal Kayacan. 2020. AU-AIR: A Multi-modal Unmanned Aerial Vehicle Dataset for Low Altitude Traffic Surveillance. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 8504–8510.

[6] M Lo Brutto and Paola Meli. 2012. Computer Vision Tools for 3D Modelling in Archaeology. *International Journal of Heritage in the Digital Era* 1, 1_suppl (2012), 1–6.

[7] Yue Cao, Mingsheng Long, Bin Liu, and Jianmin Wang. 2018. Deep Cauchy Hashing for Hamming Space Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1229–1237.

[8] Botao Chen, Xi Mu, Peng Chen, Biao Wang, Jaewan Choi, Honglyun Park, Sheng Xu, Yanlan Wu, and Hui Yang. 2021. Machine Learning-based Inversion of Water Quality Parameters in Typical Reach of the Urban River by UAV Multispectral Data. *Ecological Indicators* 133 (2021), 108434.

[9] Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: A Real-World Web Image Database from National University of Singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*. 1–9.

[10] Dario Di Palma. 2023. Retrieval-augmented recommender system: Enhancing recommender systems with large language models. In *Proceedings of the 17th ACM Conference on Recommender Systems*. 1369–1373.

[11] Cathal Gurrin, Björn Þór Jónsson, Duc Tien Dang Nguyen, Graham Healy, Jakub Lokoc, Liting Zhou, Luca Rossetto, Minh-Triet Tran, Wolfgang Hürst, Werner Bailer, et al. 2023. Introduction to The Sixth Annual Lifelog Search Challenge, LSC'23. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*. 678–679.

[12] Young-Soo Han, Jaejoon Lee, Jungmin Lee, Wonhyuk Lee, and Kyungho Lee. 2019. 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* 32, 7 (2019), 658–668.

[13] Xiangnan He, Hanwang Zhang, and Tat-Seng Chua. 2018. Recommendation Technologies for Multimedia Content.. In *ICMR*. 8.

[14] Mark J Huiskes and Michael S Lew. 2008. The Mir Flickr Retrieval Evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information retrieval*. 39–43.

[15] Longlong Jing, Elahe Vahdani, Jiaxing Tan, and Yingli Tian. 2021. Cross-Modal Center Loss for 3D Cross-Modal Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3142–3151.

[16] Omar Shahbaz Khan, Björn Þór Jónsson, Jan Zahálka, Stevan Rudinac, and Marcel Worring. 2021. Impact of Interaction Strategies on User Relevance Feedback. In *Proceedings of the 2021 International Conference on Multimedia Retrieval*. 590–598.

[17] Omar Shahbaz Khan, Jan Zahálka, and Björn Þór Jónsson. 2022. Influence of Late Fusion of High-Level Features on User Relevance Feedback for Videos. In *Proceedings of the 2nd International Workshop on Interactive Multimedia Retrieval*. 17–24.

[18] Margarita Khokhlova, Valérie Gouet-Brunet, Nathalie Abadie, and Liming Chen. 2020. Cross-Year Multi-Modal Image Retrieval Using Siamese Networks. In *2020 IEEE International Conference on Image Processing (ICIP)*. IEEE, 1–5.

[19] Chao Li, Cheng Deng, Ning Li, Wei Liu, Xinbo Gao, and Dacheng Tao. 2018. Self-Supervised Adversarial Hashing Networks for Cross-Modal Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4242–4251.

[20] Mingbao Lin, Rongrong Ji, Hong Liu, and Yongjian Wu. 2018. Supervised Online Hashing via Hadamard Codebook Learning. In *Proceedings of the 26th ACM International Conference on Multimedia*. 1635–1643.

[21] Zijia Lin, Guiguang Ding, Mingqing Hu, and Jianmin Wang. 2015. Semantics-Preserving Hashing for Cross-View Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3864–3872.

[22] Xin Liu, Zhikai Hu, Haibin Ling, and Yiu-ming Cheung. 2019. MTFH: A matrix tri-factorization hashing framework for efficient cross-modal retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 43, 3 (2019), 964–981.

[23] Jakub Lokoč, Stelios Andreadis, Werner Bailer, Aaron Duane, Cathal Gurrin, Zhixin Ma, Nicola Messina, Thao-Nhu Nguyen, Ladislav Peška, Luca Rossetto, et al. 2023. Interactive Video Retrieval in the Age of Effective Joint Embedding Deep Models: Lessons from the 11th VBS. *Multimedia Systems* 29, 6 (2023), 3481–3504.

[24] Devraj Mandal, Kunal N Chaudhury, and Soma Biswas. 2017. Generalizes Semantic Preserving Hashing for N-Label Cross-Modal Retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4076–4084.

[25] Andrew Marx, Donald McFarlane, and Ahmed Alzahrani. 2017. UAV Data for Multi-temporal Landsat Analysis of Historic Reforestation: A Case Study in Costa Rica. *International Journal of Remote Sensing* 38, 8-10 (2017), 2331–2348.

[26] Elisa Mohr, Thomas Thum, and Christian Bär. 2022. Accelerating Cardiovascular Research: Recent Advances in Translational 2D and 3D Heart Models. *European Journal of Heart Failure* 24, 10 (2022), 1778–1791.

[27] Nick Pantelidis, Maria Pegia, Damianos Galanopoulos, Konstantinos Apostolidis, Klearchos Stavrothanasopoulos, Anastasia Moumtzidou, Konstantinos Gkountakos, Ilias Gialampoukidis, Stefanos Vrochidis, Vasileios Mezaris, et al. 2024. VERGE in VBS 2024. In *International Conference on Multimedia Modeling*. Springer, 356–363.

[28] Maria Pegia, Björn Þór Jónsson, Anastasia Moumtzidou, Sotiris Diplaris, Ilias Gialampoukidis, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2024. Multimodal 3D Object Retrieval. In *International Conference on Multimedia Modeling*. Springer, 188–201.

[29] Maria Pegia, Björn Þór Jónsson, Anastasia Moumtzidou, Ilias Gialampoukidis, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2023. MuseHash: Supervised Bayesian Hashing for Multimodal Image Representation. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval*. 434–442.

[30] Maria Pegia, Ferran Agullo Lopez, Anastasia Moumtzidou, Alberto Gutierrez-Torre, Björn Þór Jónsson, Josep Lluís Berral García, Ilias Gialampoukidis, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2024. Time-Quality Tradeoff of MuseHash Query Processing Performance. In *International Conference on Multimedia Modeling*. Springer, 270–283.

[31] Maria Pegia, Anastasia Moumtzidou, Ilias Gialampoukidis, Björn Þór Jónsson, Stefanos Vrochidis, and Ioannis Kompatsiaris. 2022. BiasHash: A Bayesian Hashing Framework for Image Retrieval. In *2022 IEEE 14th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP)*. IEEE, 1–5.

[32] Renjing Pei, Jianzhuang Liu, Weimian Li, Bin Shao, Songcen Xu, Peng Dai, Juwei Lu, and Youliang Yan. 2023. CLIPPING: Distilling CLIP-Based Models with a Student Base for Video-Language Retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18983–18992.

[33] Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. 2022. Multimodal Co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion* 81 (2022), 203–239.

[34] Patricia Schöntag, David Nakath, Stefan Röhrl, and Kevin Köser. 2022. Towards Cross Domain Transfer Learning for Underwater Correspondence Search. In *International Conference on Image Analysis and Processing*. Springer, 461–472.

[35] Leon Amadeus Varga, Benjamin Kiefer, Martin Messmer, and Andreas Zell. [n. d.]. SeaDronesSee: A Maritime Benchmark for Detecting Humans in Open Water. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*.

[36] Yongxin Wang, Zhen-Duo Chen, Xin Luo, Rui Li, and Xin-Shun Xu. 2021. Fast Cross-Modal Hashing With Global and Local Similarity Embedding. *IEEE Transactions on Cybernetics* 52, 10 (2021), 10064–10077.

[37] Martin J Willemink, Wojciech A Koszek, Cailin Hardell, Jie Wu, Dominik Fleischmann, Hugh Harvey, Les R Folio, Ronald M Summers, Daniel L Rubin, and Matthew P Lungren. 2020. Preparing Medical Imaging Data for Machine Learning. *Radiology* 295, 1 (2020), 4–15.

[38] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015. 3D ShapeNets: A Deep Representation for Volumetric Shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1912–1920.

[39] Yanzhao Xie, Yu Liu, Yangtao Wang, Lianli Gao, Peng Wang, and Ke Zhou. 2020. Label-Attended Hashing for Multi-Label Image Retrieval.. In *IJCAI*. 955–962.

[40] Bo-Hyun Yun and Chang-Ho Seo. 2003. Semantic-based Information Retrieval for Content Management and Security. *Computational Intelligence* 19, 2 (2003), 87–110.

[41] Yu-Wei Zhan, Yongxin Wang, Yu Sun, Xiao-Ming Wu, Xin Luo, and Xin-Shun Xu. 2022. Discrete Online Cross-Modal Hashing. *Pattern Recognition* 122 (2022), 108262.