VISIONE: A Large-Scale Video Retrieval System with Advanced Search Functionalities

Giuseppe Amato
giuseppe.amato@isti.cnr.it
ISTI-CNR
Pisa, Italy

Paolo Bolettieri
paolo.bolettieri@isti.cnr.it
ISTI-CNR
Pisa, Italy

Fabio Carrara
fabio.carrara@isti.cnr.it
ISTI-CNR
Pisa, Italy

Fabrizio Falchi
fabbrizio.falchi@isti.cnr.it
ISTI-CNR
Pisa, Italy

Claudio Gennaro
claudio.gennaro@isti.cnr.it
ISTI-CNR
Pisa, Italy

Nicola Messina*
nicola.messina@isti.cnr.it
ISTI-CNR
Pisa, Italy

Lucia Vadicamo*
lucia.vadicamo@isti.cnr.it
ISTI-CNR
Pisa, Italy

Claudio Vairo
claudio.vairo@isti.cnr.it
ISTI-CNR
Pisa, Italy

ABSTRACT

VISIONE is a large-scale video retrieval system that integrates multiple search functionalities, including free text search, spatial color and object search, visual and semantic similarity search, and temporal search. The system leverages cutting-edge AI technology for visual analysis and advanced indexing techniques to ensure scalability. As demonstrated by its runner-up position in the 2023 Video Browser Showdown competition, VISIONE effectively integrates these capabilities to provide a comprehensive video retrieval solution. A system demo is available online, showcasing its capabilities on over 2300 hours of diverse video content (V3C1+V3C2 dataset) and 12 hours of highly redundant content (Marine dataset). The demo can be accessed at https://visione.isti.cnr.it/.

CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Users and interactive retrieval**; **Retrieval models and ranking**; *Search engine architectures and scalability*; **Multimedia and multimodal retrieval**; **Video search**.

KEYWORDS

multimedia retrieval, video search, cross-modal search, interactive system

1 INTRODUCTION

With the increasing diffusion of multimedia databases, there is today, as never before, the need to analyze, organize, and index all the produced data so that they can be easily and efficiently retrieved. The use of these systems is not only tailored to the organization of wild multimedia content uploaded every second on public social-media platforms – like Youtube or Instagram. Instead, consider large audiovisual archives owned by national televisions and updated daily with dozens of hours of unannotated content. Audiovisual documents are vital for future generations to preserve and recollect their past cultures, beliefs, and customs. The development of tools to automatically analyze and index all these contents constitutes a major achievement in the automatic content-based organization and browsing of all these audiovisual archives. In this context, Artificial Intelligence – and, in particular, Deep Learning models – defined major milestones to automatically understand multimedia content, extract information, and index data to be easily searchable, increasing the accessibility of large multimedia databases.

Despite the large global research investment in video understanding and the joint processing of different modalities, there is a lack of software tools that bundle all these technologies together for large-scale video search in an interactive and user-friendly manner. Interactivity is an important feature of search systems, where the human searcher and the search software are entangled in the same loop, collaborating to browse large video collections smartly. Benchmarking competitions, such as the Video Browser Showdown (VBS) [16] and Lifelog Search Challenge [13], are organized annually to foster research and development of such large-scale multimedia retrieval software (see [18, Table 1] for other examples of evaluation campaigns with multimedia retrieval and analysis tasks). At VBS, in particular, systems are evaluated during online search sessions where human searchers are asked to use the system to find given shots in the shortest possible time. These challenges demonstrate the large and increasing research interest in the development of large-scale interactive multimedia retrieval systems [5, 8, 15, 17, 19].

This demonstration paper presents the latest release of VISIONE [1, 2, 4–6], an interactive large-scale video search system. It recently participated in the 12th Video Browser Showdown (VBS2023) competition, where it achieved remarkable success in

*Corresponding authors

ICMR '23, June 12–15, 2023, Thessaloniki, Greece

ACM ISBN 979-8-4007-0178-8/23/06.
https://doi.org/10.1145/3591106.3592226

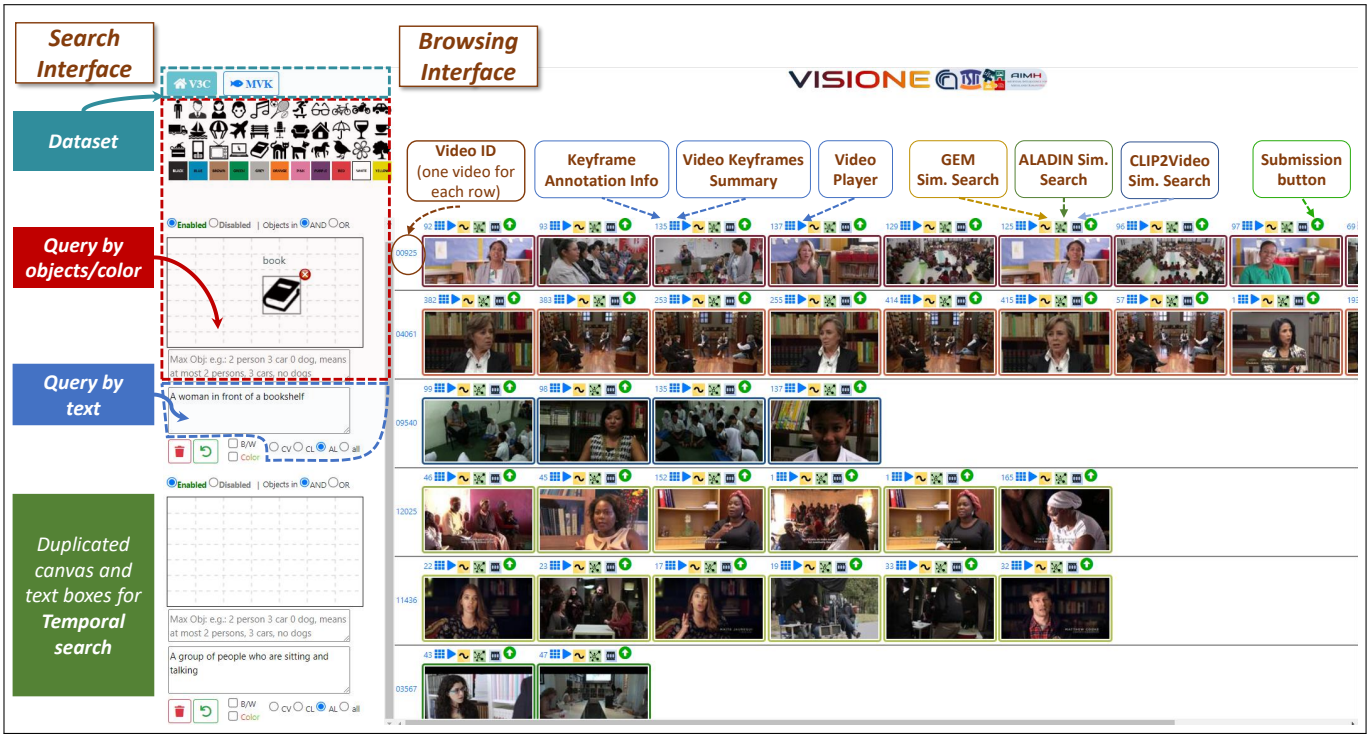

Figure 1: User Interface. Example of results using the temporal search for two video frames, one containing a "*book*" and "*a woman in front of a bookshelf*", the other "*a group of people who are sitting and talking*". Each row in the browsing interface corresponds to a video, and the first two columns contain the most relevant results according to our ranking model.

numerous tasks and came in second place in the overall leaderboard. This tool incorporates many content-based analysis tools for automatically extracting knowledge from raw shots and employs mature indexing techniques to ensure scalability. It offers several search functionalities, like searching for video shots given specific object classes and natural language prompts. VISIONE also provides various visual similarity techniques to browse results, allowing users to find keyframes similar to the selected one.

VISIONE features our recently developed cross-modal retrieval deep neural network, called ALADIN (ALign And DIstill Network) [20]. ALADIN generates easily indexable and fixed-length features lying in a common visual-textual space. This capability helps bridge the gap between different modalities of digital media and user-generated queries, allowing quick and accurate media retrieval.

A system demo and a video showcasing its capabilities are available online at https://visione.isti.cnr.it/ and https://youtu.be/iiecKRDv05g, respectively. The demo allows exploring and searching over 2300 hours of diverse video content (V3C dataset [23]) and 12 hours of highly redundant content (Marine dataset [24]).

2 THE VISIONE SYSTEM

VISIONE provides multiple search options for users to retrieve a specific video segment. These search functionalities include text-based and visual-based queries, as well as the ability to search two temporally close video frames. In particular, VISIONE supports free text search, spatial color and object search, visual similarity search, and semantic similarity search. We report an example of the user interface in Figure 1 and the system design in Figure 2.

2.1 Objects and Colors

VISIONE enables video frame search by placing plaparticular object classes and colors in a canvas, where the location of the specific object and/or color within the frame can be specified. Furthermore, it is also possible to constrain the maximum number of instances of a particular object class (e.g., *4 persons* and *1 dog*).

To implement object-based search, we employed three separated object detectors (VfNet [26], Mask R-CNN [14], Faster R-CNN [12]) trained respectively on the COCO, LVIS, and Open Images V4 datasets, each having its own set of classes. We mapped these classes using a semi-automatic process to obtain a unified final set of 1,460 classes that we organized into a hierarchy using WordNet. The hierarchy is used to expand class labels during indexing and query runtime. We released the final list of classes and the corresponding hierarchy in [3]. For the color annotations, we employed two chip-based color naming techniques [9, 25].

2.2 Text-to-Image and Text-to-Video Retrieval

VISIONE supports different technologies for searching videos through natural language descriptions of a desired scene. Specifically, we employed two CLIP-based models: CLIP [21] trained on the

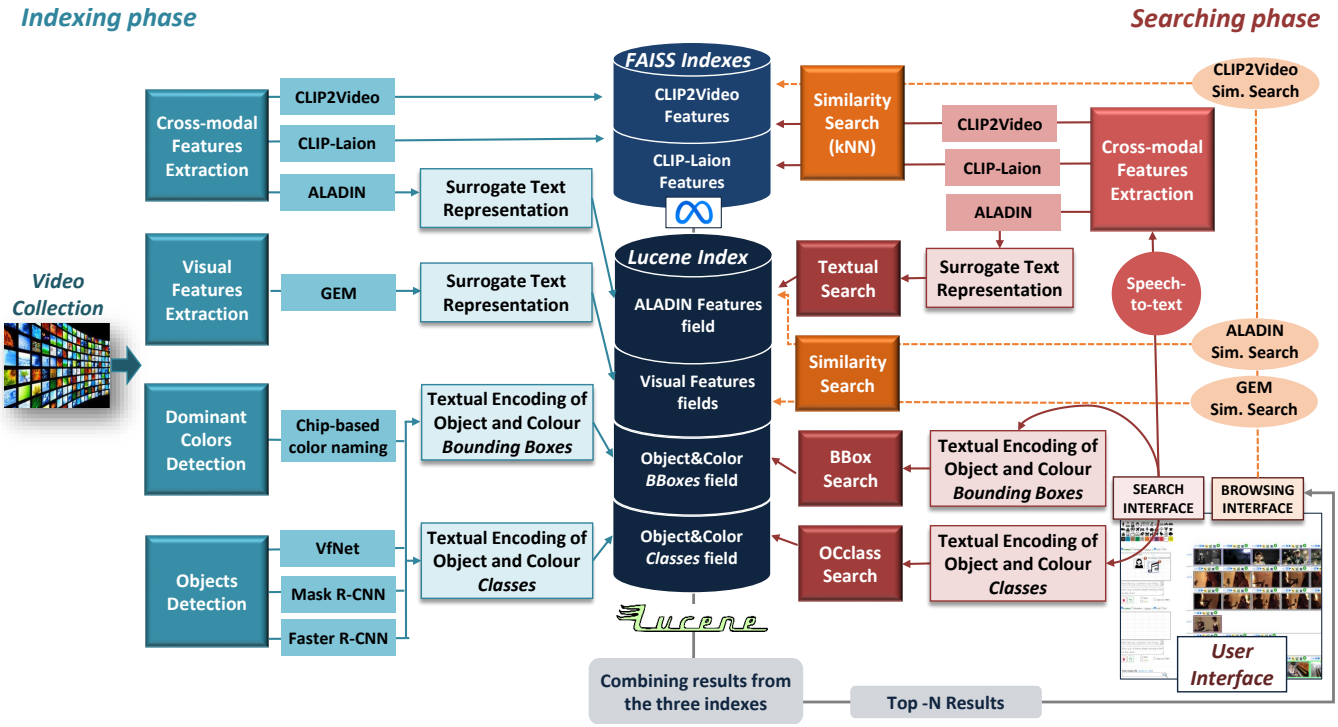

Figure 2: VISIONE System Architecture

LAION dataset[1], and CLIP2Video [11], able to perform, respectively, retrieval of keyframes and shots from natural language prompts. The advantage of CLIP2Video over plain CLIP is its ability to understand the temporal dimension, which is useful when the textual prompt includes long-lasting actions or a sequence of events.

In VISIONE, we also implemented a novel cross-modal retrieval model called ALADIN (ALign And DIstill Network). The network, as we described in [20], first generates high-quality scores by precisely aligning images and texts using the features from a large pre-trained vision language transformer. Then, it uses the scores produced by its cross-modal alignment head – very effective yet quite computationally expensive – to train a shared embedding space, allowing for an efficient and effective inference by performing a kNN search in this learned space. Specifically, the network employs a learning-to-rank loss to distill the relevance scores and train the matching head comprising a low-dimensional (768-d) cross-modal embedding space. Empirically, we found that ALADIN performs competitively with state-of-the-art vision-language Transformers while being approximately 90 times faster during inference. ALADIN is different from CLIP-based models, as (i) it employs a different training mechanism, (ii) features an entirely different visual backbone, which extracts features from an object detector instead of directly from pixels, and (iii) it is trained using overall a much fewer amount of image-text pairs. In practice, we observed that ALADIN is often complementary to CLIP since, in many cases, only one of the two methods finds the exact keyframe in the top results. Thus, in VISIONE, we designed an algorithm for combining results derived from ALADIN and the CLIP-based models.

2.3 Visual and Semantic Similarity

VISIONE allows users to query by example through the results shown in the interface. It supports both *visual* and *semantic* similarity searches. For visual similarity search, the user can use an image as the query to search for video keyframes visually similar to it (e.g., similar background, same building, etc). In semantic similarity search, an image can be used to retrieve video keyframes or video clips that are semantically similar to it (e.g., scenes with similar descriptions). GEM features [22] are used for visual similarity search, while CLIP2Video [11] and ALADIN [20] are used for searching semantically similar video clips and video keyframes, respectively.

2.4 Temporal Queries

VISIONE supports temporal queries allowing the user to specify two different queries, which we will refer to as a and b. A temporal quantization approach is used for searching videos that contain one keyframe satisfying query a and another satisfying b. The time is divided into intervals of T seconds (e.g., $T = 3$), and the results of both queries are independently processed to retain a single representative result (the one with the highest score) for each time interval and for each query. Result pairs (a_i, b_j) that come from the same video and have a temporal distance smaller than 12 seconds are displayed to the user as results. Temporal quantization is also utilized to present a limited number of result pairs from the same video, where only the pair with the highest aggregated score in a

[1]https://huggingface.co/laion/CLIP-ViT-L-14-laion2B-s32B-b82K

specific time interval is considered. Figure 1 shows an example of a temporal query.

2.5 Indexing

As shown in Figure 2, we employed two different indexes to store and perform similarity search on the extracted visual features and the detected objects and colors. Specifically, we employed Apache Lucene[2]and FAISS[3]. The need for two indexes is motivated by their different functionalities and implementations. We originally decided to rely on Lucene as it is disk-based and scales very well to billions of documents, releasing the need for strong main memory requirements as in-memory indexes like FAISS. Lucene is commonly used for text-based search in collections of long unstructured text documents. However, it can be employed for data encoded in the form of text, like quantized colors and object classes, together with their quantized 2D coordinates in the frame, as in our case [2]. We developed a family of techniques called Surrogate Text Representations (STRs) [7, 10] to index feature vectors extracted from neural networks in Lucene. STRs enable dense features to be transformed into sparse term frequencies from an appropriate codebook, preserving the dot product between the obtained textual representation and the original dense feature as much as possible.

While we found the STR approach to work well on many features like GEM for visual similarity or ALADIN for text-based search, CLIP-based features demonstrated some major problems with this encoding technique. In particular, we noticed the mean cosine similarity between the query text and the top nearest neighboring images for the CLIP2Video (Figure 3a) and CLIP LAION (Figure 3b) features is considerably lower than the one from ALADIN features (Figure 3c). This may happen if element-wise products underlying the dot-product computation have a negative sign, which implies that there could be a lot of mixed-sign factors. This is a bad scenario for the STR representation, given that the CReLU operation at the core of the STR method zeroes out the contribution from mixed-signed factors. Therefore, for the CLIP2Video features, the approximated cosine similarity computed in the STR representation badly approximates the original one. For these reasons, for the CLIP-based features, we instead relied on the FAISS index, using an exact search and an 8-bit scalar quantization to reduce the index size in memory[4]. Despite the exact search, with the in-memory quantized index, the search over the full V3C1 + V3C2 shots takes only a few milliseconds at a cost of much bigger memory utilization. Although according to Figures 3a and 3b the STR representation for image-to-image search should not have the same problems as text-to-image search, we also relied on FAISS for the semantic similarity search for these CLIP-based features, for ease of implementation.

3 CONCLUSIONS

In this work, we presented VISIONE, a large-scale and interactive video search tool. Inspired by state-of-the-art techniques in cross-modal analysis and image-video understanding, it implements many user-friendly tools for searching among large video collections.

[2]https://lucene.apache.org/
[3]https://github.com/facebookresearch/faiss
[4]We leave the investigation of a STR technique that is suitable for indexing this type of dense vector to future work.

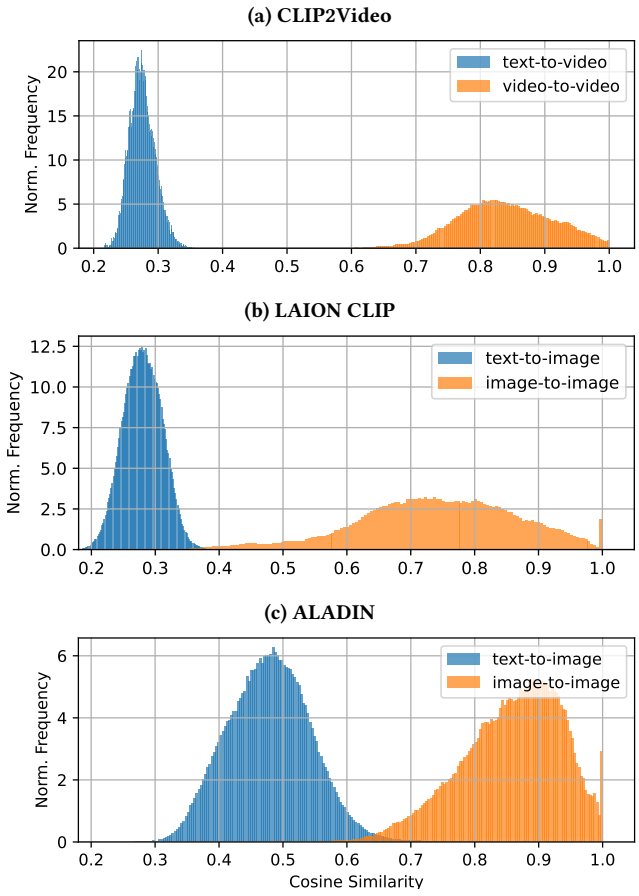

Figure 3: Distribution of cosine similarities between nearest neighbors for different cross-modal models.

Specifically, it implements object queries by placing the desired objects or colors in a canvas, it allows video searching by specifying natural language descriptions of desired keyframes or shots, and it supports temporal queries for finding consecutive specific events.

Future work on the VISIONE system should focus on unifying the indexing methods to reduce memory requirements and accessing dynamic knowledge bases to improve the retrieval of visual named entities, such as famous persons or buildings. Furthermore, we plan to increase its interactiveness by developing advanced tools for suggesting textual query changes based on the current result set.

ACKNOWLEDGMENTS

This work was partially funded by AI4Media - A European Excellence Centre for Media, Society and Democracy (EC, H2020 n. 951911) and National Centre for HPC, Big Data and Quantum Computing – HPC (CUP B93C22000620006).

REFERENCES

[1] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Franca Debole, Fabrizio Falchi, Claudio Gennaro, Lucia Vadicamo, and Claudio Vairo. 2019. VISIONE at VBS2019. In *MultiMedia Modeling*. Springer, 591–596.

[2] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Franca Debole, Fabrizio Falchi, Claudio Gennaro, Lucia Vadicamo, and Claudio Vairo. 2021. The VISIONE video search system: exploiting off-the-shelf text search engines for large-scale video retrieval. *Journal of Imaging* 7, 5 (2021), 76.

[3] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Lucia Vadicamo, and Claudio Vairo. 2022. *COCO, LVIS, Open Images V4 classes mapping*. https://doi.org/10.5281/zenodo.7194300

[4] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Lucia Vadicamo, and Claudio Vairo. 2022. VISIONE at Video Browser Showdown 2022. In *MultiMedia Modeling*. Springer International Publishing, Cham, 543–548.

[5] Giuseppe Amato, Paolo Bolettieri, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Lucia Vadicamo, and Claudio Vairo. 2023. VISIONE at Video Browser Showdown 2023. In *MultiMedia Modeling*, Duc-Tien Dang-Nguyen, Cathal Gurrin, Martha Larson, Alan F. Smeaton, Stevan Rudinac, Minh-Son Dao, Christoph Trattner, and Phoebe Chen (Eds.). Springer International Publishing, Cham, 615–621.

[6] Giuseppe Amato, Paolo Bolettieri, Fabrizio Falchi, Claudio Gennaro, Nicola Messina, Lucia Vadicamo, and Claudio Vairo. 2021. VISIONE at Video Browser Showdown 2021. In *International Conference on Multimedia Modeling*. Springer, 473–478.

[7] Giuseppe Amato, Fabio Carrara, Fabrizio Falchi, Claudio Gennaro, and Lucia Vadicamo. 2019. Large-scale instance-level image retrieval. *Information Processing & Management* (2019), 102100.

[8] Stelios Andreadis, Anastasia Moumtzidou, Damianos Galanopoulos, Nick Pantelidis, Konstantinos Apostolidis, Despoina Touska, Konstantinos Gkountakos, Maria Pegia, Ilias Gialampoukidis, Stefanos Vrochidis, Vasileios Mezaris, and Ioannis Kompatsiaris. 2022. VERGE in VBS 2022. In *MultiMedia Modeling*. Springer International Publishing, Cham, 530–536.

[9] Robert Benavente, Maria Vanrell, and Ramon Baldrich. 2008. Parametric fuzzy sets for automatic color naming. *JOSA A* 25, 10 (2008), 2582–2593.

[10] Fabio Carrara, Lucia Vadicamo, Claudio Gennaro, and Giuseppe Amato. 2022. Approximate Nearest Neighbor Search on Standard Search Engines. In *Similarity Search and Applications*, Tomáš Skopal, Fabrizio Falchi, Jakub Lokoč, Maria Luisa Sapino, Ilaria Bartolini, and Marco Patella (Eds.). Springer International Publishing, Cham, 214–221.

[11] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. 2021. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097* (2021).

[12] Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 1440–1448.

[13] Cathal Gurrin, Liting Zhou, Graham Healy, Björn Þór Jónsson, Duc-Tien Dang-Nguyen, Jakub Lokoč, Minh-Triet Tran, Wolfgang Hürst, Luca Rossetto, and Klaus Schöffmann. 2022. Introduction to the Fifth Annual Lifelog Search Challenge, LSC'22. In *International Conference on Multimedia Retrieval (ICMR'22). ACM*.

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969.

[15] Silvan Heller, Rahel Arnold, Ralph Gasser, Viktor Gsteiger, Mahnaz Parian-Scherb, Luca Rossetto, Loris Sauter, Florian Spiess, and Heiko Schuldt. 2022. Multimodal interactive video retrieval with temporal queries. In *MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part II*. Springer, 493–498.

[16] Silvan Heller, Viktor Gsteiger, Werner Bailer, Cathal Gurrin, Björn Þór Jónsson, Jakub Lokoč, Andreas Leibetseder, František Mejzlík, Ladislav Peška, Luca Rossetto, et al. 2022. Interactive video retrieval evaluation at a distance: comparing sixteen interactive video search systems in a remote setting at the 10th Video Browser Showdown. *International Journal of Multimedia Information Retrieval* 11, 1 (2022), 1–18.

[17] Nico Hezel, Konstantin Schall, Klaus Jung, and Kai Uwe Barthel. 2022. Efficient Search and Browsing of Large-Scale Video Collections with Vibro. In *MultiMedia Modeling*. Springer International Publishing, Cham, 487–492.

[18] Jakub Lokoč, Werner Bailer, Kai Uwe Barthel, Cathal Gurrin, Silvan Heller, Björn Þór Jónsson, Ladislav Peška, Luca Rossetto, Klaus Schoeffmann, Lucia Vadicamo, et al. 2022. A task category space for user-centric comparative multimedia search evaluations. In *MultiMedia Modeling: 28th International Conference, MMM 2022, Phu Quoc, Vietnam, June 6–10, 2022, Proceedings, Part I*. Springer, 193–204.

[19] Jakub Lokoč, František Mejzlík, Tomáš Souček, Patrik Dokoupil, and Ladislav Peška. 2022. Video Search with Context-Aware Ranker and Relevance Feedback. In *MultiMedia Modeling*. Springer International Publishing, Cham, 505–510.

[20] Nicola Messina, Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, Giuseppe Amato, and Rita Cucchiara. 2022. ALADIN: Distilling Fine-grained Alignment Scores for Efficient Image-Text Matching and Retrieval. *arXiv preprint arXiv:2207.14757* (2022).

[21] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

[22] J. Revaud, J. Almazan, R.S. Rezende, and C.R. de Souza. 2019. Learning with Average Precision: Training Image Retrieval with a Listwise Loss. In *International Conference on Computer Vision*. IEEE, 5106–5115.

[23] Luca Rossetto, Heiko Schuldt, George Awad, and Asad A Butt. 2019. V3C–a research video collection. In *International Conference on Multimedia Modeling*. Springer, 349–360.

[24] Quang-Trung Truong, Tuan-Anh Vu, Tan-Sang Ha, Jakub Lokoč, Yue Him Wong Tim, Ajay Joneja, and Sai-Kit Yeung. 2023. Marine Video Kit: A New Marine Video Dataset for Content-based Analysis and Retrieval. In *MultiMedia Modeling - 29th International Conference, MMM 2023, Bergen, Norway, January 9-12, 2023*.

[25] Joost Van De Weijer, Cordelia Schmid, Jakob Verbeek, and Diane Larlus. 2009. Learning color names for real-world applications. *IEEE Transactions on Image Processing* 18, 7 (2009), 1512–1523.

[26] Haoyang Zhang, Ying Wang, Feras Dayoub, and Niko Sunderhauf. 2021. VarifocalNet: An IoU-aware Dense Object Detector. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.