LifeGraph 4 – Lifelog Retrieval using Multimodal Knowledge Graphs and Vision-Language Models

Luca Rossetto
Department of Informatics
University of Zurich
Zurich, Switzerland
rossetto@ifi.uzh.ch

Athina Kyriakou
Department of Informatics
University of Zurich
Zurich, Switzerland
kyriakou@ifi.uzh.ch

Svenja Lange
Department of Informatics
University of Zurich
Zurich, Switzerland
lange@ifi.uzh.ch

Florian Ruosch
Department of Informatics
University of Zurich
Zurich, Switzerland
ruosch@ifi.uzh.ch

Ruijie Wang
Department of Informatics
University of Zurich
Zurich, Switzerland
ruijie@ifi.uzh.ch

Kathrin Wardatzky
Department of Informatics
University of Zurich
Zurich, Switzerland
wardatzky@ifi.uzh.ch

Abraham Bernstein
Department of Informatics
University of Zurich
Zurich, Switzerland
bernstein@ifi.uzh.ch

ABSTRACT

In the scope of the 7th Lifelog Search Challenge (LSC'24), we present the 4th iteration of LifeGraph, a multimodal knowledge-graph approach with data augmentations using Vision-Language Models (VLM). We extend the LifeGraph model presented in former LSC challenges by event-based clustering using temporal and spatial relations as well as information extracted from descriptions of Lifelog image captions produced by VLMs.

CCS CONCEPTS

• **Information systems** → *Users and interactive retrieval*; *Specialized information retrieval*; **Multimedia and multimodal retrieval**.

KEYWORDS

Lifelogging, Lifelog Search Challenge, Multimodal Knowledge Graphs, Graph-based Retrieval, Multi-modal Retrieval, Vision-Language Models

ACM Reference Format:

Luca Rossetto, Athina Kyriakou, Svenja Lange, Florian Ruosch, Ruijie Wang, Kathrin Wardatzky, and Abraham Bernstein. 2024. LifeGraph 4 – Lifelog Retrieval using Multimodal Knowledge Graphs and Vision-Language Models. In *The 7th Annual ACM Lifelog Search Challenge (LSC '24), June 10, 2024, Phuket, Thailand*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3643489.3661127

1 INTRODUCTION

Lifelogs are inherently multi-modal collections of autobiographical data with various internal and external relations and are, therefore, well suited to be represented in a multi-modal graph structure. Relying only on information that can easily be represented in a structured form, however, limits the available retrieval options. Modern semantic embedding methods, combined with classical vector space retrieval techniques, can overcome such limitations by enabling search for concepts and scenarios that are not preemptively annotated or even annotatable.

In this paper, we present our contribution to the 2024 Lifelog Search Challenge: LifeGraph 4. Since the dataset used in 2024 [4] is the same as in the previous two years [3], we build upon the 2023 version of our graph-based approach [12] by re-using some of the fundamental structures while removing some previous structures that turned out not to be beneficial. We augment the graph with information extracted from the images using a state-of-the-art Vision-Language Model and introduce some modified querying and result presentation schemes.

The remainder of this paper is structured as follows: Section 2 describes our data cleaning and pre-processing steps and graph construction method. Section 3 then gives an overview of the available querying mechanisms, and Section 4 describes the interaction modes available to a user. In Section 5, we show the results of some preliminary experiments based on queries from 2023 before we conclude the paper in Section 6.

LSC '24, June 10, 2024, Phuket, Thailand

ACM ISBN 979-8-4007-0550-2/24/06
https://doi.org/10.1145/3643489.3661127

2 DATA PRE-PROCESSING

Our graph construction process involves several data cleaning and feature extraction steps, which are outlined in the following.

2.1 Data Cleaning

As indicated in previous implementations [1, 16], a large portion of the dataset consists of blurry, obstructed, or dark images, which are not helpful for the challenge's retrieval tasks. Filtering out these unhelpful images could ameliorate our system's performance. However, due to the redaction of the dataset's images for privacy preservation [4], characterizing an image as unhelpful is not a trivial task.

To identify unhelpful images, we first extracted SIFT features [10] for all images, as previously proposed by Tran et al. [16]. We filtered out the images for which no SIFT features were detected. Secondly, we evaluated the blurriness of the remaining images using focus measures. We filtered out blurry images based on empirical thresholds.

2.2 Usage of Vision-Language Models

We used two Vision-Language Models (VLM) to extract information from the images in the dataset: BLIP-2 [8] and LLaVA [9]. Due to their models' inherent natures, we employed different approaches for the two. Both were run as local instances and prompted for each image.

For BLIP-2, we used three questions to extract information from the images. The first was aimed at identifying the objects present ("What are the objects in this image?"). The second was to extract the environment as natural language text ("What is the environment seen in this image?"). The third sought to determine the location or context of the image ("Where was this picture taken?"). Examples of the BLIP-2 output are shown in Table 1 relating to the images in Figures 1, 2, and 3.

Furthermore, we employed LLaVA to generate descriptions of the images by prompting it with "Describe the image!". This resulted in detailed text such as the examples shown in Table 1, again relating to images in Figure 1, 2, and 3.

Taking a closer look at the descriptions provided by LLaVA and BLIP-2 (see Table 1), we can find several significant relations to the test queries for the first two example images, while the descriptions generated for the third image are lacking important information. Concerning the image in Figure 1 and the first query about drinks on a rooftop in Bangkok, keywords such as "rooftoop", "glass", "drink" (by LLaVA), and "night" (by BLIP-2) are helpful for finding the corresponding lifelog images. However, the outputs of the two VLM's cannot account for neither completeness, nor accuracy. For example, LLaVA describes the first image to feature "a person standing", while BLIP-2 for example describes a cigarette and the picture as taken "at the top of the world, the world's tallest building, the chicago tower".

It is also remarkable that the descriptions provided by LLaVA change over sequences of similar images, which are characteristic to lifelog data. For example, in the sequence of images showing the drinks on a rooftop in Bangkok, LLaVA changes from describing a person "standing" to "sitting" among others, while there are no explicit indicators in the images.

Considering the image in Figure 2, LlaVA mislocates the dog on the right side - and closely behind the man, while BLIP-2 again creates a very specific description of the location as "the garden of a house in the village of clonmel, county tipperary". The third image (see Figure 3) is rather tricky. The information about the model train mentioned in the test query can mainly be extracted by OCR. While both VLM's correctly describe the scene as showing the interior of a store, both fail to find more relevant information for the test query. Unfortunately, LLaVA misdescribes the image as showing a grocery store, "aisles filled with food items", "several people shopping", "several TVs mounted", and as featuring a dining area, too. BLIP-2 does not provide much of information, but also misdescribes "a box of cereals".

In the final step, we extracted the concepts from the answers of the VLM's. This was done using Core NLP's [11] open information extraction pipeline. Then, we mapped the extracted concepts to Wikidata's[1] Q-identifiers through entity linking. Furthermore, we also kept the generated data to use as a fallback option in full-text search.

[1]https://www.wikidata.org/

2.3 Embedding Methods for Similarity Search

To enable text-based querying, we use two independent embedding models. The first one, which was already used in our 2023 participation [12], uses an OpenCLIP [2] model trained on the LAION-5B [14] dataset.

The second one uses the VLM-generated textual descriptions described in the previous section. We employ OpenAI Text Embeddings[2] to retrieve images for given natural language queries based on the text descriptions of the images generated by LLaVA. Specifically, we adopt the model text-embedding-3-small[3] to encode the text descriptions of all images into contextual embeddings of the size 1536. Then, given a natural language query, we encode the query with the same model and rank all images according to the cosine similarity of their embeddings to the embedding of the query.

2.4 Event Detection for Temporal Query Handling

The Lifelog data possess a sequential structure that can be harnessed for temporal query handling. Yet, for temporal retrieval to be effective, the images need to be organized into semantically related sequences, previously referred to as "events" [1, 16]. We segment the available images into events based on the images' spatial, temporal, and visual information.

2.4.1 Temporal. We primarily use the provided temporal metadata associated with each image [4].

2.4.2 Spatial. Similarly to the previous year's participation [12], to determine the lifelogger's physical location, we use the provided metadata, and we infer further spatial information based on visual input. In particular, we use the "latitude", "longitude", and "semantic name" columns of the metadata table. We refer to this information as *provided semantic location*. Additionally, we infer spatial information based on visual input. To do so, we query Wikidata for the closest physical entity with a spatial position to any log entry. We refer to this information as *inferred semantic location*.

2.4.3 Visual. We identify the visual concepts present on each image using a combination of the identified (a) SIFT features (Section 2.1) and (b) VGG16 features [15], as well as the computed embeddings (c) from the OpenCLIP model trained on the LAION-5B dataset and (d) the VLM-generated textual descriptions (Section 2.3).

2.4.4 Event identification algorithm. To identify events, we process the images sequentially. Each image is compared with its immediately preceding one based on the identified visual features and using cosine distance. Two images are considered to belong to the same event if their cosine distance is below an empirically identified threshold. Each event is given a unique identifier. It is annotated with its start and end time and provided and inferred locations.

[2]https://platform.openai.com/docs/guides/embeddings
[3]https://openai.com/blog/new-embedding-models-and-api-updates

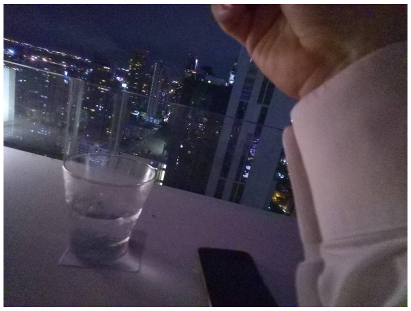

Figure 1: Image 20190915_140614_000

Figure 2: Image 20191225_133531_000

Figure 3: Image 20190622_081018_000

Table 1: Test queries and VLM output corresponding to the lifelog images in Figure 1, 2, and 3

	Figure 1	Figure 2	Figure 3
Test query	Drinks on top of the Bangkok. Taking a drink on a rooftop bar at night in Bangkok. It was on the same day that I flew into Bangkok.	I remember a man in a blue coat walking a dog in the countryside in Ireland on a sunny afternoon.	When did I buy that model train? I remember it was a marklin brand train and I bought it at the weekend. Jer convinced me to buy it when having coffee and I bought it immediately after coffee. It was in June 2019.
LLaVA	The image features a person standing on a balcony or rooftop, with a city skyline visible in the background. The person is holding a glass of water, possibly enjoying a drink while taking in the view. The glass is placed on a table nearby. In addition to the person and the glass, there is a cell phone on the table, possibly belonging to the person or someone else who was there earlier. The scene captures a moment of relaxation and appreciation of the city's beauty.	The image features a man walking his dog on a leash down a dirt road. The man is wearing a blue jacket and is positioned towards the left side of the scene. The dog is located on the right side of the man, following closely behind him. The dirt road appears to be a pathway through a wooded area, with trees surrounding the scene. The overall atmosphere suggests a peaceful and leisurely walk for both the man and his dog.	The image depicts a large grocery store with aisles filled with various food items. The store has a clean and organized appearance, making it easy for customers to find what they are looking for. There are several people shopping in the store, with some browsing the aisles and others standing near the cash registers. In addition to the food items, there are several TVs mounted on the walls throughout the store, likely providing information or entertainment for the shoppers. The store also features a dining area with multiple dining tables and chairs, possibly for customers to enjoy their meals or take a break from shopping.
BLIP-2	a) A glass of water and a cigarette b) The environment is seen in this image is a city at night c) at the top of the world, the world's tallest building, the chicago tower	a) a man walking his dog on a country road b) a man walking his dog on a country road c) it was taken in the garden of a house in the village of clonmel, county tipperary	a) A box of cereal b) a supermarket c) in a supermarket

Test queries are extracted from the LSC'23 archive, LLava output for the prompt "Describe the image!", and BLIP-2 output corresponding to prompts: a) "What are the objects in this image?", b) "What is the environment seen in this image?", and c) "Where was this picture taken?".

2.5 Multimodal Graph Construction

As in previous iterations, the graph structure is organized around the images provided in the dataset since they form the unit of retrieval for two out of the three task types. Each image is related to the entities and feature vectors extracted from them, as well as to the metadata provided with the dataset. In addition, we re-use some of the clustering introduced in our 2023 participation [12] to group-related images. These groupings are then used for result aggregation. Additionally, higher-order structures for larger time intervals (i.e., days, months, years) are used for efficient filtering.

As in our previous system version, the graph is stored in our custom MediaGraph Store[4] that is capable of storing and querying multimodal graphs, including their media content.

3 QUERYING

This section provides a brief overview of the querying and interaction mechanism offered by our graph-based approach.

[4] https://github.com/lucaro/MeGraS

3.1 Unstructured Data

The initial query is most commonly done using the features described in Section 2.3. Both embedding methods generate a query vector from a text query, which is then used to perform a kNN search to determine relevant images. The two embedding methods can be used in isolation, or their results can be combined using score-fusion. For each image, the graph structure will also be queried for its temporal context.

3.2 Relevance Feedback

We apply a relevance feedback approach inspired by Khan et al. [6, 7] to include a human signal in the querying process. After the initial query provides a set of images, the user can label each image as relevant or not relevant. A classifier then uses this information to determine the embedding vectors of the images that are closest to the relevant images and furthest to the non-relevant ones within the embedding space.

3.3 Structured Data

In addition to queries based on vector information, the structured data in the graph can also be used for querying. These queries can include all types of relationships captured by the graph and be either used as an initial query or as a late filter. When used as an initial query, the graph is queried directly for all images that are associated with the relevant properties. When used as a late filter, an existing result set is used, and all contained elements that do not match the filter criteria are removed.

3.4 Temporal Query Handling

Inspired by previous LSC systems [1, 17] and based on our definition of events (Section 2.4), we enable the temporal search of events in close vicinity to a main event. Any retrieved event or other aggregate of results can be used as a query to obtain information about what happened in close temporal proximity before or after the selected event. Additionally, queries can be constructed with a temporal notion from the outset by combining multiple sub-queries with a first-this-then-that temporal semantic.

4 USER INTERACTION

Similar to its predecessor, LifeGraph 4 uses a minimalist user interface that devotes as much screen estate as possible to result-exploration. Textual and structural queries can be entered at the top of the screen, analogously to [13]. After the results for a query have been returned, this top bar also offers controls over how results are to be grouped for display.

In addition to the option of narrowing down the set of retrieved images by providing relevance feedback as described in Section 3.2, we implement the option to re-rank the results with the goal of increasing the diversity of the images within the retrieved set. We use a maximal marginal relevance recommendation approach to minimize the number of very similar images being seen at the same time, aiming to increase the efficiency of exploration. As different use cases and query results might require different re-ranking strategies, we provide the option to adapt the degree of diversity dynamically.

Table 2: Test queries as used in LSC'23 and the ranking of the 1st correct image by our system. Showing results for both known-item search queries (top) and ad-hoc search queries (bottom).

Query*	Embeddings	OpenCLIP
Drinks on top of the Bangkok...	1	1
I remember that there was a man...	1	1
When did I buy that model train...	28,279	17,666
I was getting an eye test after...	93	1,010
Having lunch with Dermot, who...	18,883	29
I remember a man in a blue coat...	1	1
At a hungry lunchtime, I was eating...	14	30
There was a man in the front row...	22	1
I am never eating BBQ'd oysters...	1	3
Waiting at Dublin airport to collect...	20	1,112
Find examples of when I was eating...	14	5
Find examples of when I was taking...	3	83
Find examples of when I was trying...	6	6
Find examples of me reading a menu...	1	1
Find examples when I'm in the car...	3	1
Find examples of when I was in...	2	40
Find examples of me using an...	1	1
Find examples of when I was wearing...	257	2
Find examples of me taking a picture...	1	1
Find examples of when I was looking...	2	1

* Refer to the archive of LSC'23 for complete queries.[5]

Retrieved elements can also be used to expand the current result set by requesting the temporal context of any retrieved event group. Similarly, current result sets can also be filtered without re-issuing a query in order to narrow down the results and hide elements later deemed to be irrelevant from view. This functionality is primarily used for structured metadata to hide results that do not originate from a specific location or time interval. Since, during certain task types, more information is revealed during the task, it can be more efficient to narrow down already retrieved results rather than reissuing a new query with additional constraints.

5 PRELIMINARY EXPERIMENTS

We conducted preliminary experiments to examine the performance of our system when using the embeddings computed by OpenAI Embeddings and OpenCLIP [5]. The embeddings of OpenAI Embeddings are computed based on the LLaVA descriptions, as introduced in Section 2.2 and Section 2.3. The embeddings of OpenCLIP are computed based on "xlm-roberta-base-ViT-B-32."[6]

There were 20 queries in LSC 2023[7] that required the retrieval of specific images. We use them as test queries in a zero-shot setup—the system was not optimized for this challenge. In LSC 2023, ground-truth images were provided for 9 test queries. For the rest 11 test queries, we use the answers from participants that were judged to be correct in the challenge. We aim to evaluate if

[5]https://github.com/lucaro/LSC-Archive/blob/main/2023/LSC23.json
[6]https://github.com/mlfoundations/open_clip
[7]http://lifelogsearch.org/lsc/2023/index.html

the images ranked at the top according to the cosine similarity of computed embeddings are indeed correct answers. The ranking of the first correct answer for each test query is reported in Table 2, where we can observe that the results are very promising. A correct image is returned in the top 20 candidates for 15 (75%) and 13 (65%) queries based on OpenAI Embeddings and OpenCLIP, respectively. Especially, the 1st-ranked image is directly a correct answer for 7 (35%) and 9 (45%) queries based on the two embedding methods.

6 CONCLUSION

In this paper, we presented an overview of our contribution to the 2024 Lifelog Search Challenge. Our graph-based approach builds upon our contribution from last year by extending it with additional information extracted using a Vision-Language Model. Preliminary experiments indicate that the new features are capable of solving a large fraction of the known-item search and ad-hoc search queries from 2023 while complementing the retrieval performance of the existing CLIP-based functionality.

ACKNOWLEDGMENTS

This work was partially funded by the Digital Society Initiative of the University of Zurich, the University Research Priority Program "Dynamics of Healthy Aging" at the University of Zurich, and the Swiss National Science Foundation through Projects "CrowdAlytics" (Grant Number 184994) and "MediaGraph" (Grant Number 202125).

REFERENCES

[1] Naushad Alam, Yvette Graham, and Cathal Gurrin. 2021. Memento: A Prototype Lifelog Search Engine for LSC'21. In *Proceedings of the 4th Annual on Lifelog Search Challenge* (Taipei, Taiwan) *(LSC '21)*. Association for Computing Machinery, New York, NY, USA, 53–58. https://doi.org/10.1145/3463948.3469069

[2] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2022. Reproducible scaling laws for contrastive language-image learning. *CoRR* abs/2212.07143 (2022). https://doi.org/10.48550/arXiv.2212.07143 arXiv:2212.07143

[3] Cathal Gurrin, Björn Þór Jónsson, Duc-Tien Dang-Nguyen, Graham Healy, Jakub Lokoc, Liting Zhou, Luca Rossetto, Minh-Triet Tran, Wolfgang Hürst, Werner Bailer, and Klaus Schoeffmann. 2023. Introduction to the Sixth Annual Lifelog Search Challenge, LSC'23. In *Proceedings of the 2023 ACM International Conference on Multimedia Retrieval, ICMR 2023, Thessaloniki, Greece, June 12-15, 2023*. ACM, 678–679. https://doi.org/10.1145/3591106.3592304

[4] Cathal Gurrin, Liting Zhou, Graham Healy, Werner Bailer, Duc-Tien Dang-Nguyen, Steve Hodges, Björn Þór, Jakub Lokoc, Luca Rossetto, Minh-Triet Tran, and Klaus Schoeffmann. 2024. Introduction to the Seventh Annual Lifelog Search Challenge, LSC'24. In *Proceedings of the 2024 International Conference on Multimedia Retrieval* (Phuket, Thailand) *(ICMR '24)*. Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3652583.3658891

[5] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. *OpenCLIP*. https://doi.org/10.5281/zenodo.5143773 If you use this software, please cite it as below..

[6] Omar Shahbaz Khan, Aaron Duane, Björn Þór Jónsson, Jan Zahálka, Stevan Rudinac, and Marcel Worring. 2021. Exquisitor at the Lifelog Search Challenge 2021: Relationships Between Semantic Classifiers. In *Proceedings of the 4th Annual on Lifelog Search Challenge* (Taipei, Taiwan) *(LSC '21)*. Association for Computing Machinery, New York, NY, USA, 3–6. https://doi.org/10.1145/3463948.3469255

[7] Omar Shahbaz Khan, Björn Þór Jónsson, Jan Zahálka, Stevan Rudinac, and Marcel Worring. 2019. Exquisitor at the Lifelog Search Challenge 2019. In *Proceedings of the ACM Workshop on Lifelog Search Challenge* (Ottawa ON, Canada) *(LSC '19)*. Association for Computing Machinery, New York, NY, USA, 7–11. https://doi.org/10.1145/3326460.3329156

[8] Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA (Proceedings of Machine Learning Research, Vol. 202)*, Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (Eds.). PMLR, 19730–19742. https://proceedings.mlr.press/v202/li23q.html

[9] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (Eds.). http://papers.nips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html

[10] David G Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the seventh IEEE international conference on computer vision*, Vol. 2. Ieee, 1150–1157.

[11] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Rose Finkel, Steven Bethard, and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, ACL 2014, June 22-27, 2014, Baltimore, MD, USA, System Demonstrations*. The Association for Computer Linguistics, 55–60. https://doi.org/10.3115/V1/P14-5010

[12] Luca Rossetto, Oana Inel, Svenja Lange, Florian Ruosch, Ruijie Wang, and Abraham Bernstein. 2023. Multi-Mode Clustering for Graph-Based Lifelog Retrieval. In *Proceedings of the 6th Annual ACM Lifelog Search Challenge, LSC 2023, Thessaloniki, Greece, June 12-15, 2023*. ACM, 36–40. https://doi.org/10.1145/3592573.3593102

[13] Loris Sauter, Heiko Schuldt, Raphael Waltenspül, and Luca Rossetto. 2023. Novice-Friendly Text-based Video Search with vitrivr. In *20th International Conference on Content-based Multimedia Indexing, CBMI 2023, Orleans, France, September 20-22, 2023*. ACM, 163–167. https://doi.org/10.1145/3617233.3617262

[14] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. *CoRR* abs/2210.08402 (2022). https://doi.org/10.48550/arXiv.2210.08402 arXiv:2210.08402

[15] Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

[16] Ly-Duyen Tran, Manh-Duy Nguyen, Binh Nguyen, Hyowon Lee, Liting Zhou, and Cathal Gurrin. 2022. E-Myscéal: Embedding-based Interactive Lifelog Retrieval System for LSC'22. In *Proceedings of the 5th Annual on Lifelog Search Challenge* (Newark, NJ, USA) *(LSC '22)*. Association for Computing Machinery, New York, NY, USA, 32–37. https://doi.org/10.1145/3512729.3533012

[17] Ly-Duyen Tran, Manh-Duy Nguyen, Binh Nguyen, Hyowon Lee, Liting Zhou, and Cathal Gurrin. 2022. E-Myscéal: Embedding-based Interactive Lifelog Retrieval System for LSC'22. In *Proceedings of the 5th Annual on Lifelog Search Challenge* (Newark, NJ, USA) *(LSC '22)*. Association for Computing Machinery, New York, NY, USA, 32–37. https://doi.org/10.1145/3512729.3533012