Multimedia Retrieval in and for XR

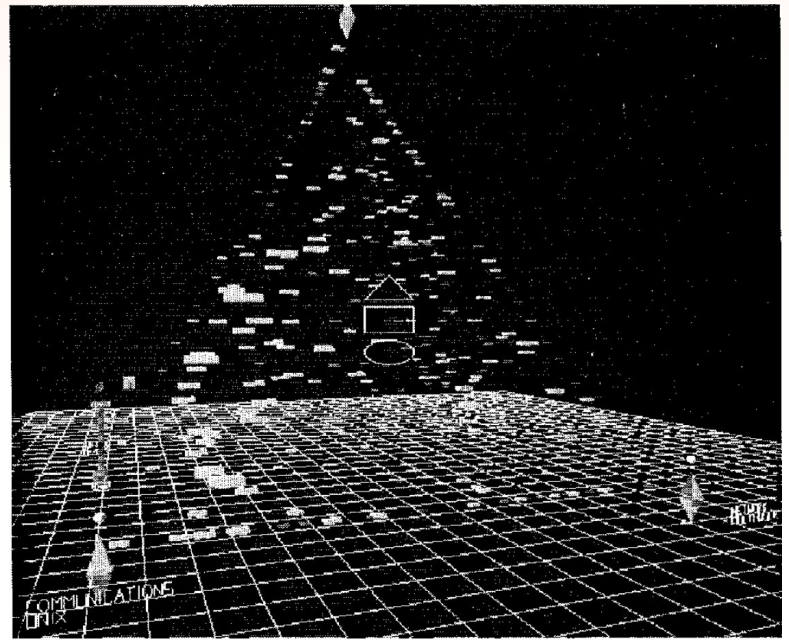
Maria Pegia, Sotiris Diplaris, Stefanos Vrochidis (CERTH-ITI),
Heiko Schuldt, **Florian Spiess**, Rahel Arnold (University of Basel),
Werner Bailer (JOANNEUM RESEARCH)

ICMR, Phuket, June 2024

Searching in VR is not a new idea …

User interfaces will be graphical. (In large data constructs, the user will be able to select an interface paradigm that suits his or her particular interests or skills; for example, a person may opt for an airplane control panel interface, not a mouse and keyboard. The control-deck paradigm makes it easier for the user to steer a search when flying over and through data).

S. Arnold, 1990 [A90]



VR-VIBE, 1995 [B95]

But today we have mobile devices …

Virtual reality for palmtop computers (1993!) [F93]

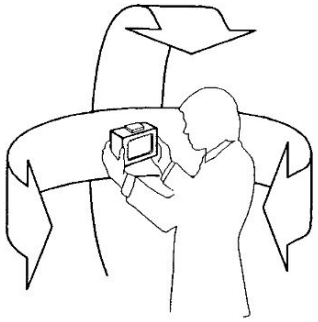


Fig. 11. Extending the virtual-reality workspace around the user in a donut shape.

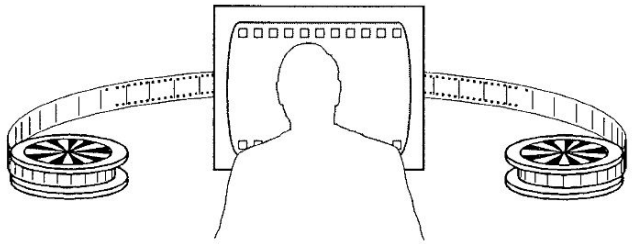


Fig 12. Film reel metaphor in which data placed on the surface of the film can be viewed in detail by the conventional computer monitor. The palmtop unit serves as a means for advancing or rewinding the film and to preview portions of the film not currently shown on the main monitor.

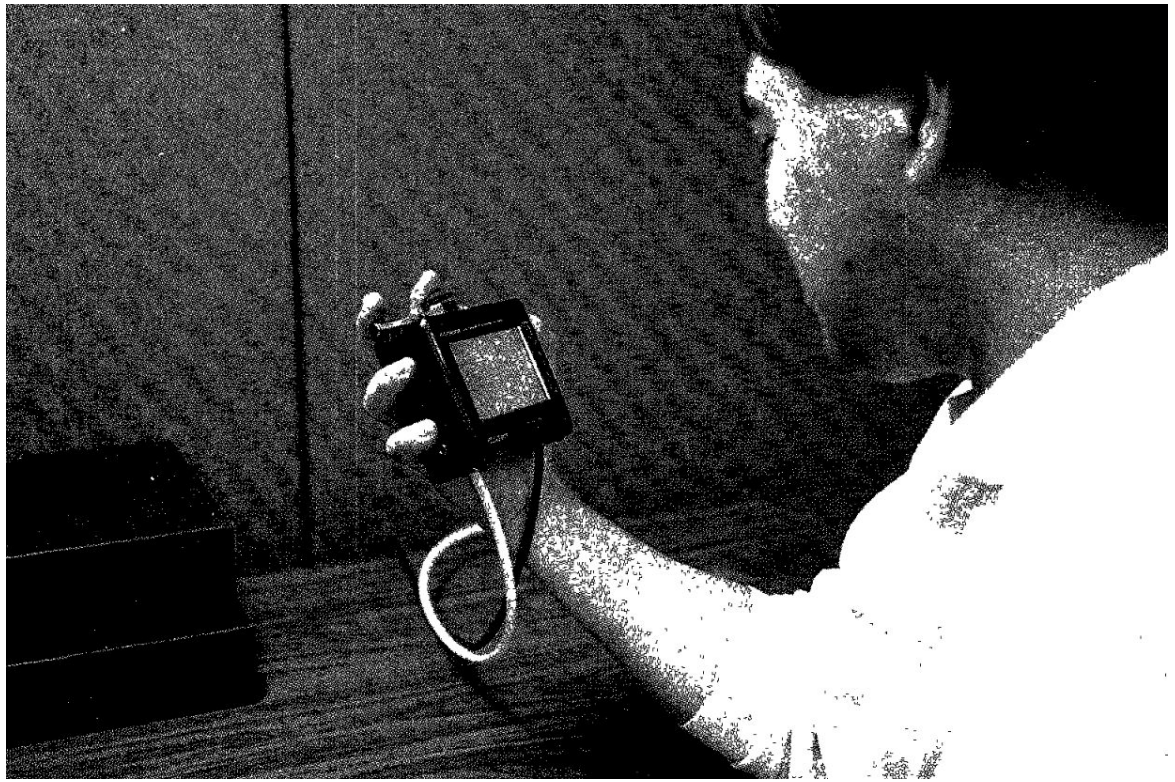
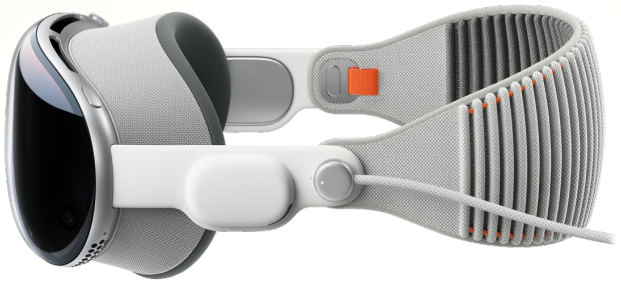


Image source: [F93]

… But today we have mobile devices

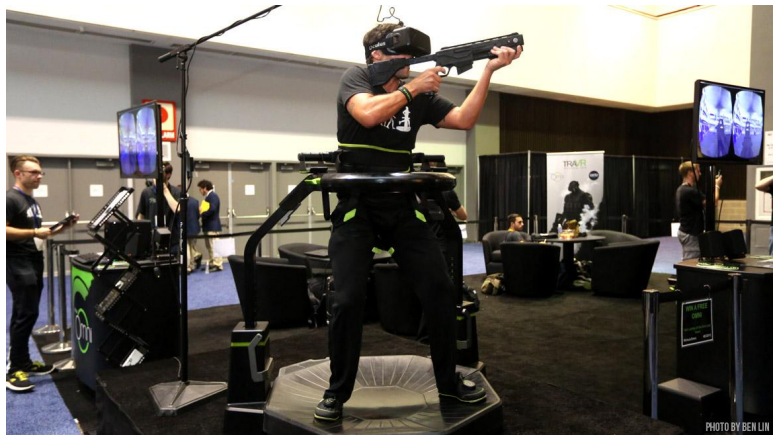
Virtual Reality in 2024

- Has become commodity
- Example: Apple Vision pro



source: https://www.apple.com/

- Advances mainly because of gaming industry
- Objective: fully immersive experience
- Example: VR goggles combined with VR treadmill



source: https://www.goodworklabs.com/

So, are we already there?

- Search and retrieval in XR seems a natural paradigm, what are the obstacles to making it work?
- Which interaction paradigms work for which type of queries?
- How do we know how well an XR-based search systems performs in realistic task settings?
- Creating XR experiences still needs a lot of manual work to create scenes and assets – what can we do about it?

Contents of this tutorial

- Introduction
- Search and exploration in XR
- Retrieval for creating XR experiences
  - 2D Asset retrieval for reconstruction
  - 3D object retrieval
- Evaluating XR search and exploration
- Q&A
- ***Break at 15:30***
- Hands-on: vitrivr-VR
- Q&A

Context: XReco project

B2B Platform for

- Creating
- Managing
- Sharing

- 3D Assets/ Scenes and

- XR Experiences

https://xreco.eu/

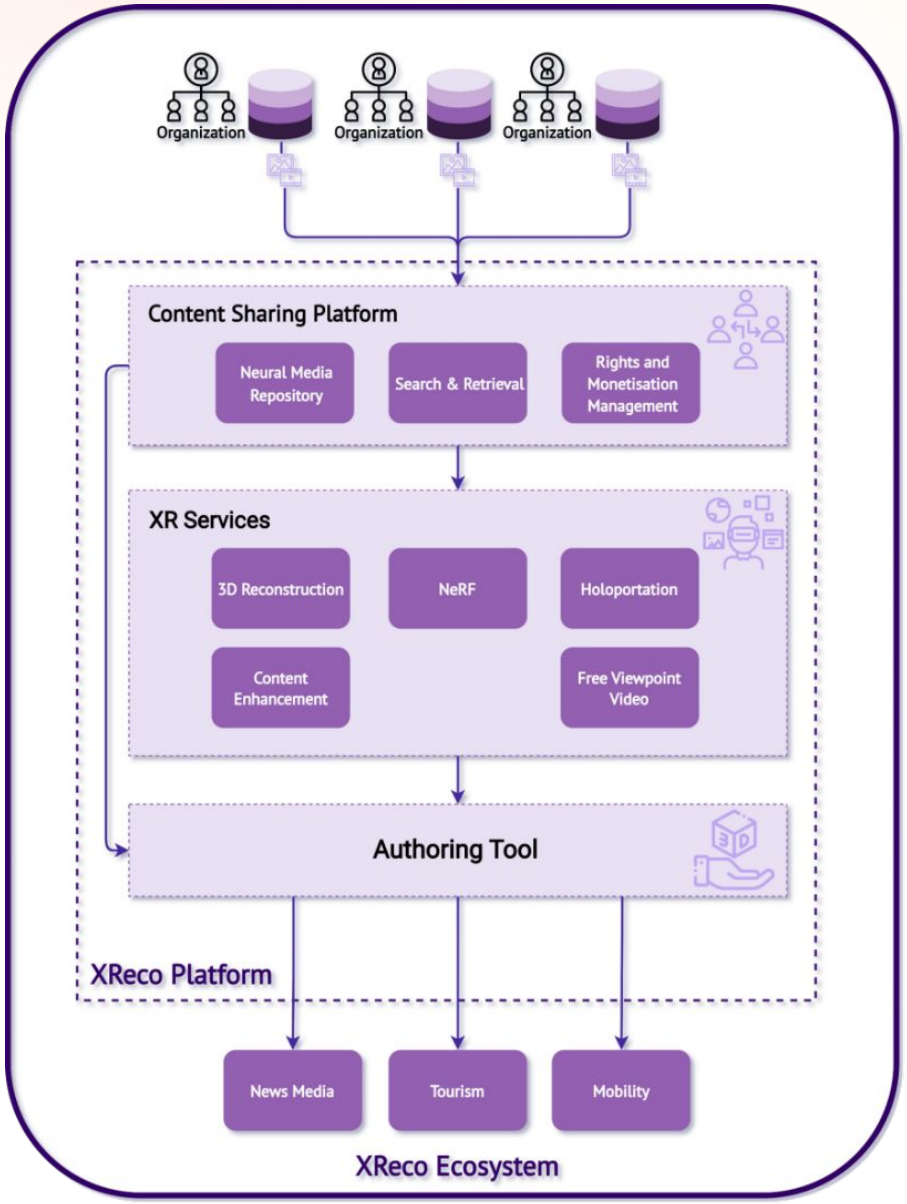
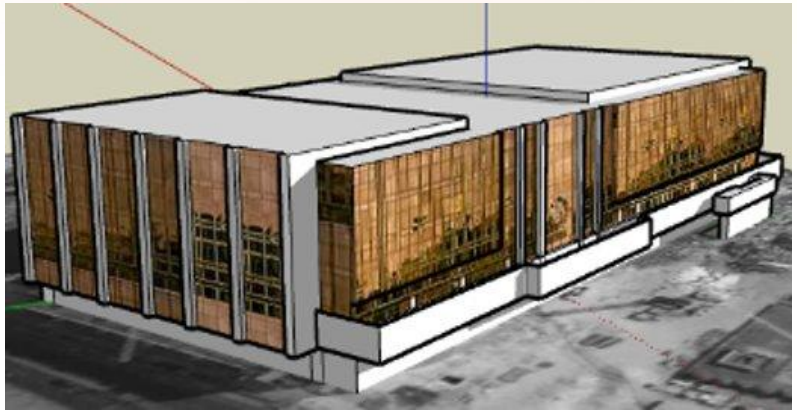


Image source: XReco project

Demonstrators

virtual production

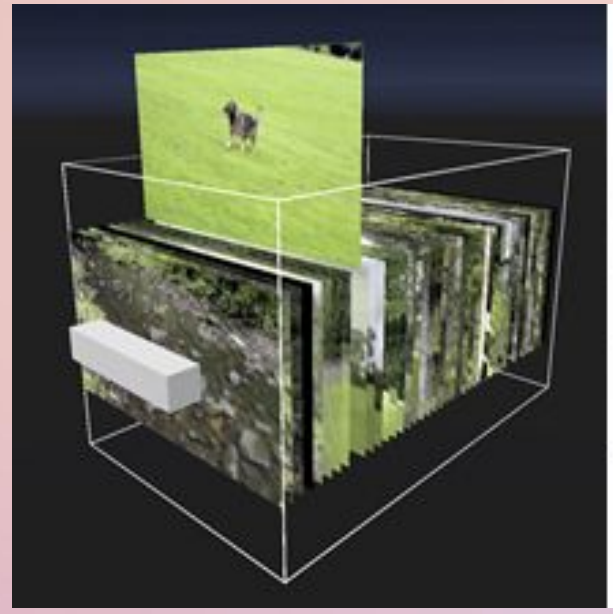


AR table-top experiences

in-car entertainment

mobile AR

Search and Exploration in XR



Query Formulation: Bridging Several Gaps

General challenge: how to properly express a user's information need

- **Sensory gap:** how a real-world object is represented digitally (e.g., 2D representation of a 3D object)
- **Semantic gap**: low-level representation vs. high-level semantic concepts (e.g., digital representation of a piece of music vs. mood)
- **Expressive gap**: how well can a user express the content of an object (e.g., the mood of a video)

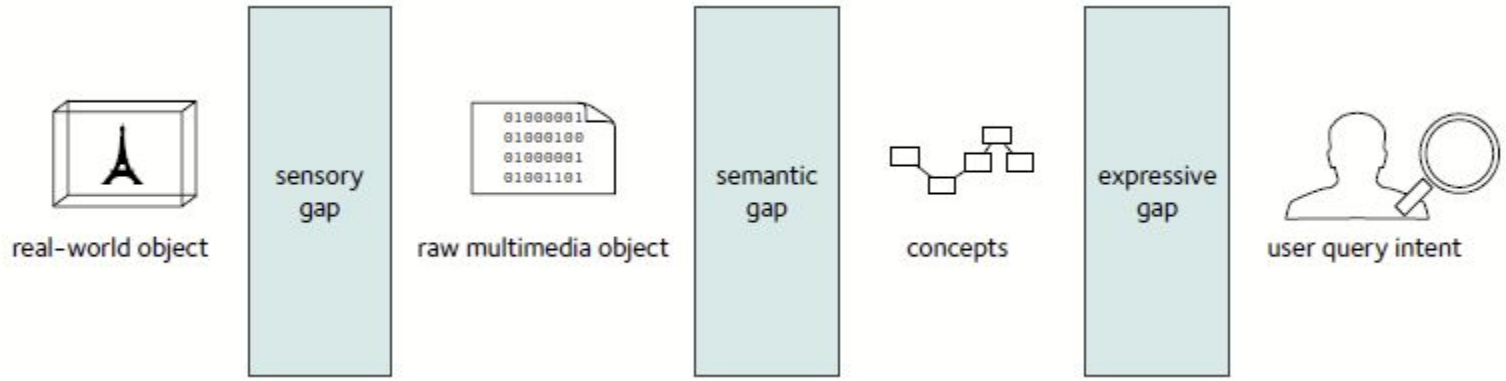


Image source: [G18]

Query Formulation: Search Paradigms

In addition to “traditional” search paradigms and modalities,
XR provides novel opportunities (and also challenges) to query formulation

Image source: Midjourney

Search Paradigms: Keyword Queries

Basis: automated concept detection (co-embedding)

Challenge in XR: how to express a keyword search

- Speech-to-text interface
- Virtual keyboard

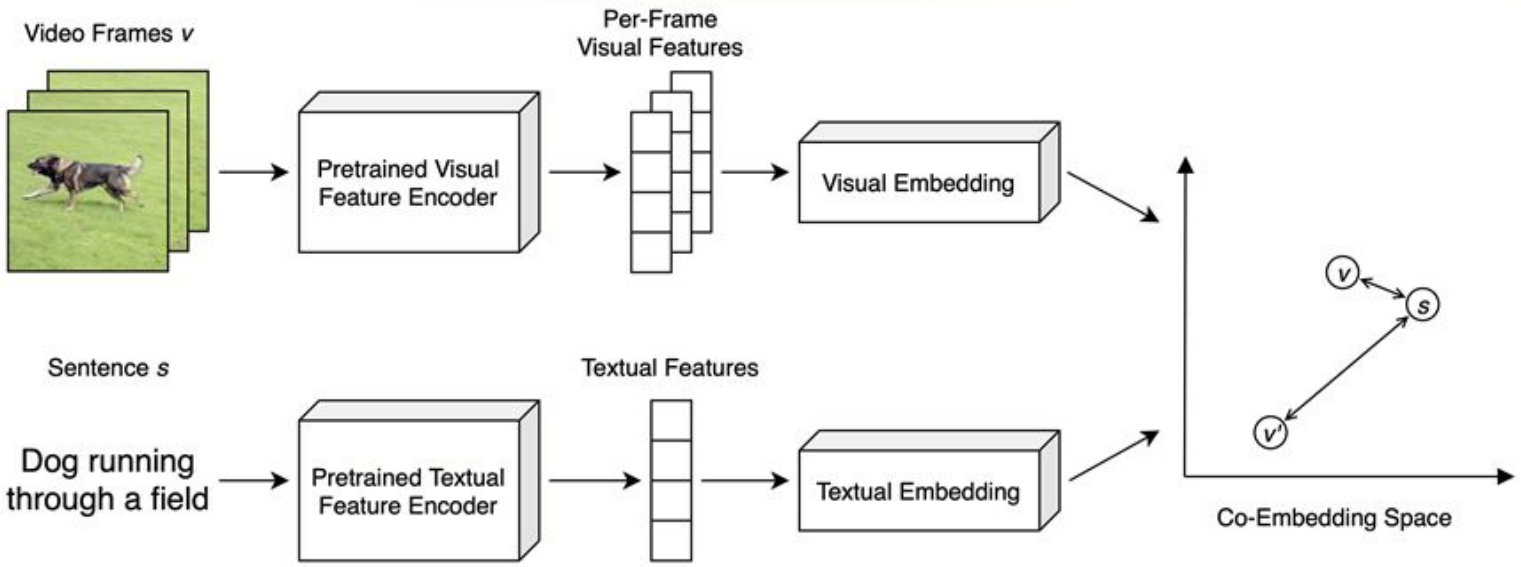


Image source: [S23a]

Search Paradigms: Query-by-Example (QbE)

Basis: object in the same modality as result

- But: search often degenerates to “more of the same”

Challenge in XR: how to get a good query object

Often used for query refinement

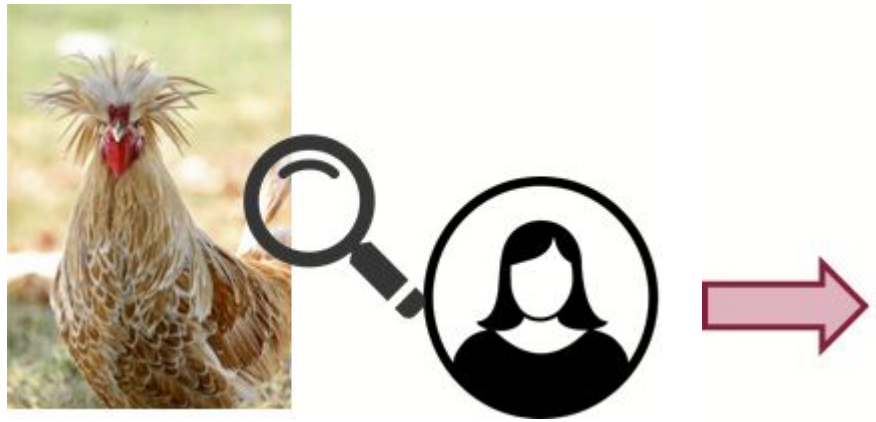


Image sources:

https://x.com/Greenpeace/status/316866068368986112

https://www.flickr.com/photos/thebigwranch12/773973981

https://www.pexels.com/photo/close-up-of-polish-hen-16776579/

https://unsplash.com/de/s/fotos/Polnisches-Hiuhn

https://www.pinterest.com/pin/striking-bufflaced-polish-rooster-definitely-a-chicken-demanding-notice--200269514666252676/

Search Paradigms: Query-by-Sketch[ing] (QbS)

Basis: rough, low-fidelity representation of the query

- But: needs some artistic skills

Challenge in XR: how to create a sketch for a non-2D object

- virtual sculpting
- humming

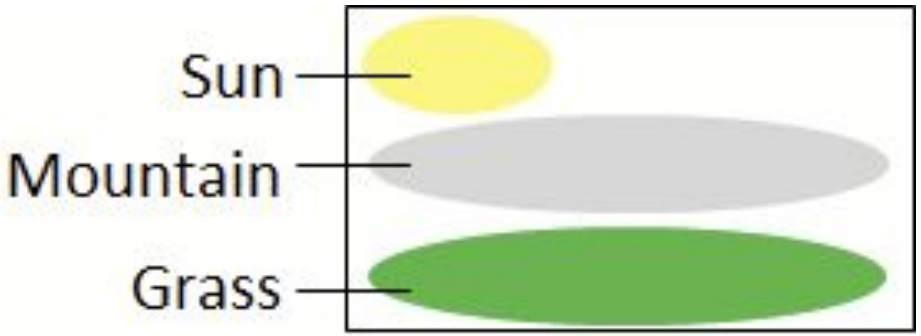
Image source: https://nycpix.wordpress.com/2011/10/23/take-home-a-street-portrait/

Search Paradigms: Query-by-SemanticSketch

Basis: combination of concept detection (keyword search) and sketching

- Sketch: region within an object, enriched with semantic label
- Addresses spatial relationships between objects

Challenge in XR: expression of spatial relationships



Search Paradigms: Query-by-BodyPose

Basis: specification of angles of a human being's joints

Challenges in XR: actually easier than in traditional user interactions

But: human anatomy is the limit

Use cases:

- Gaming
- Sports

Similarly:

- Query-by-Gesture
- Query-by-Dance

Image source: https://www.self.com/story/advanced-yoga-poses

Search Paradigms: Query-by-Time/Location

Basis: specification of object metadata (where and/or when an object was created)

Challenges in XR: how to seamlessly integrate the result(s) in the real-world scene

Use cases:

- Tourism
- Cultural heritage

Image sources: L. Sauter, Univ. Basel

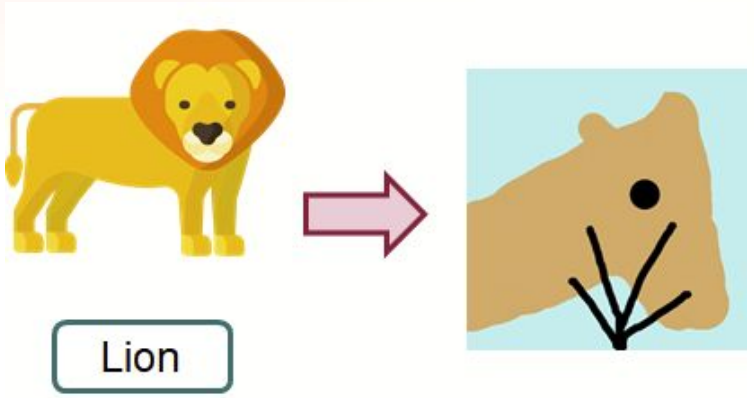
Search Paradigms: Query-by-Motion

Basis: specification of the evolution of objects over time (for non-static media)

Can be done by

- providing static queries at different points in time
- explicitly specifying motion patterns

Challenges in XR: how to specify temporal dependencies and/or trajectories



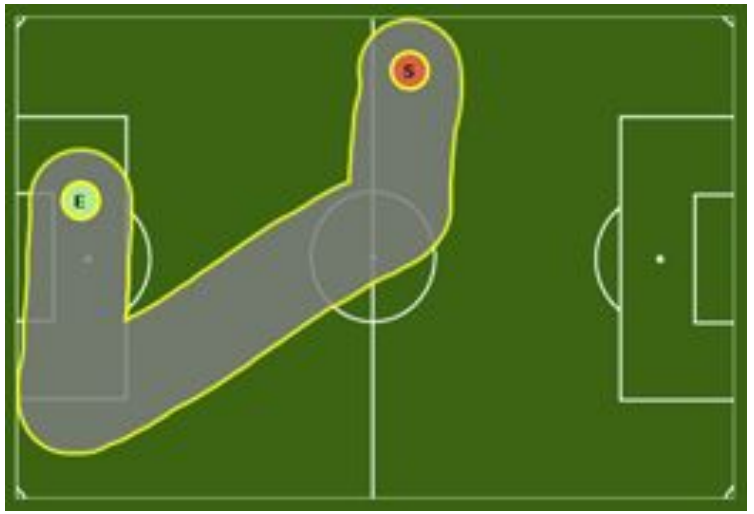
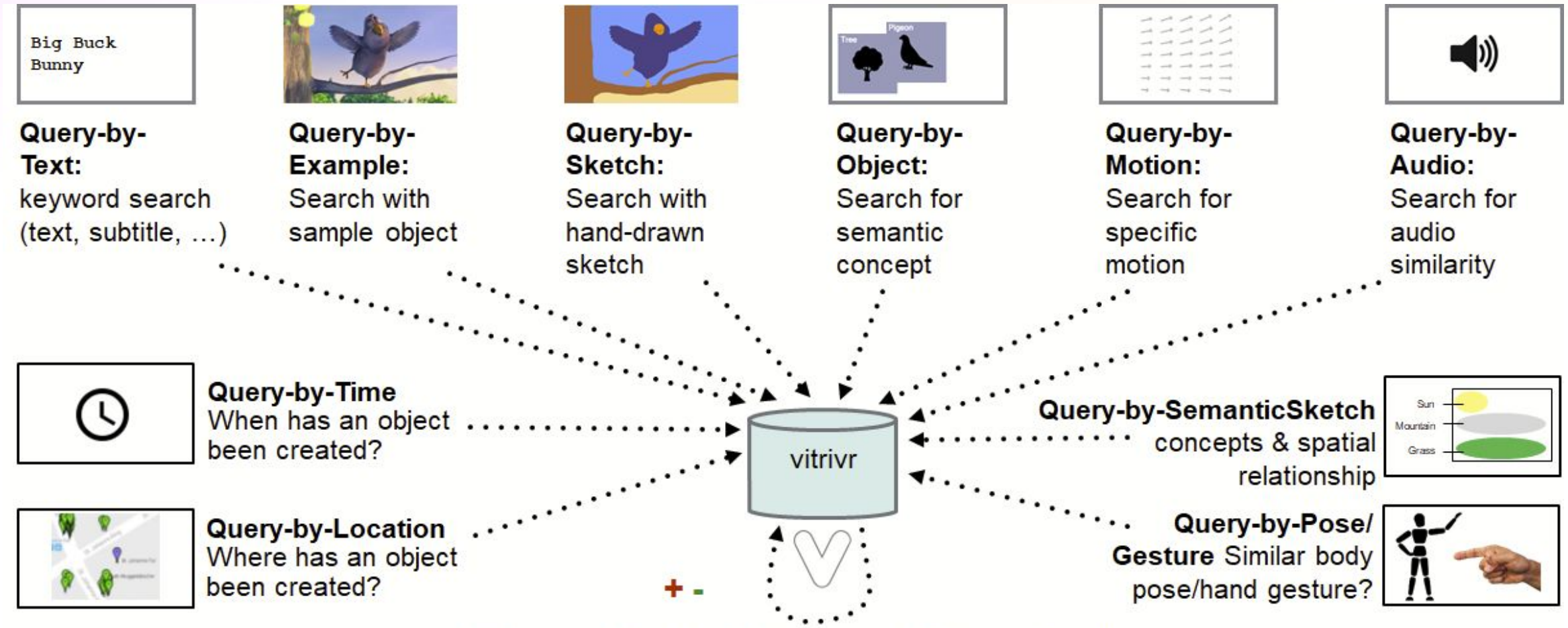
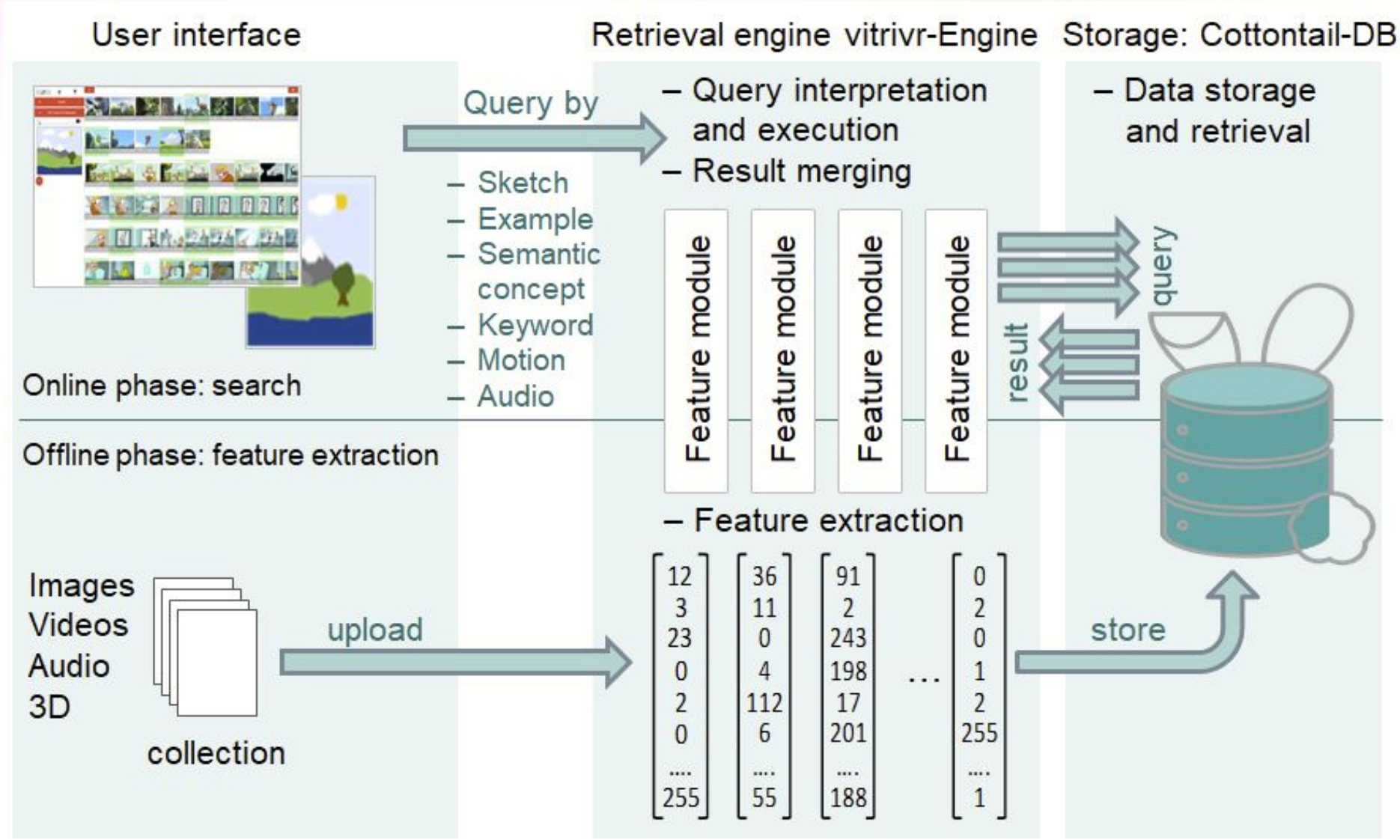


Image sources: S. Heller, I. Al Kabary, Univ. Basel

Search Paradigms: vitrivr Example



vitrivr: Architectural Overview



Search in XR

Query formulation in XR:
Real time object detection (e.g., using Yolo [C24])

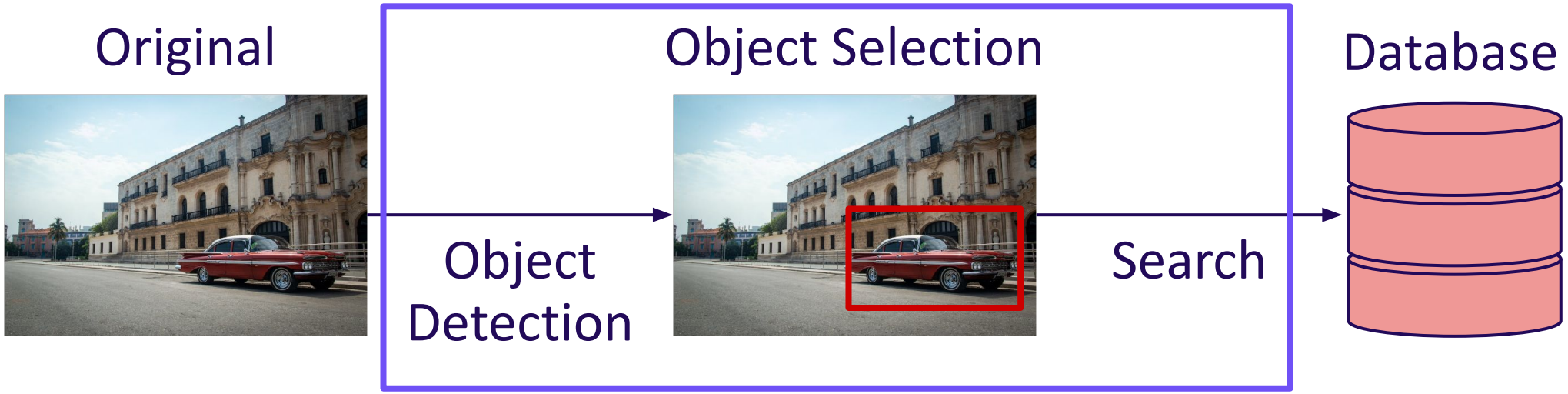
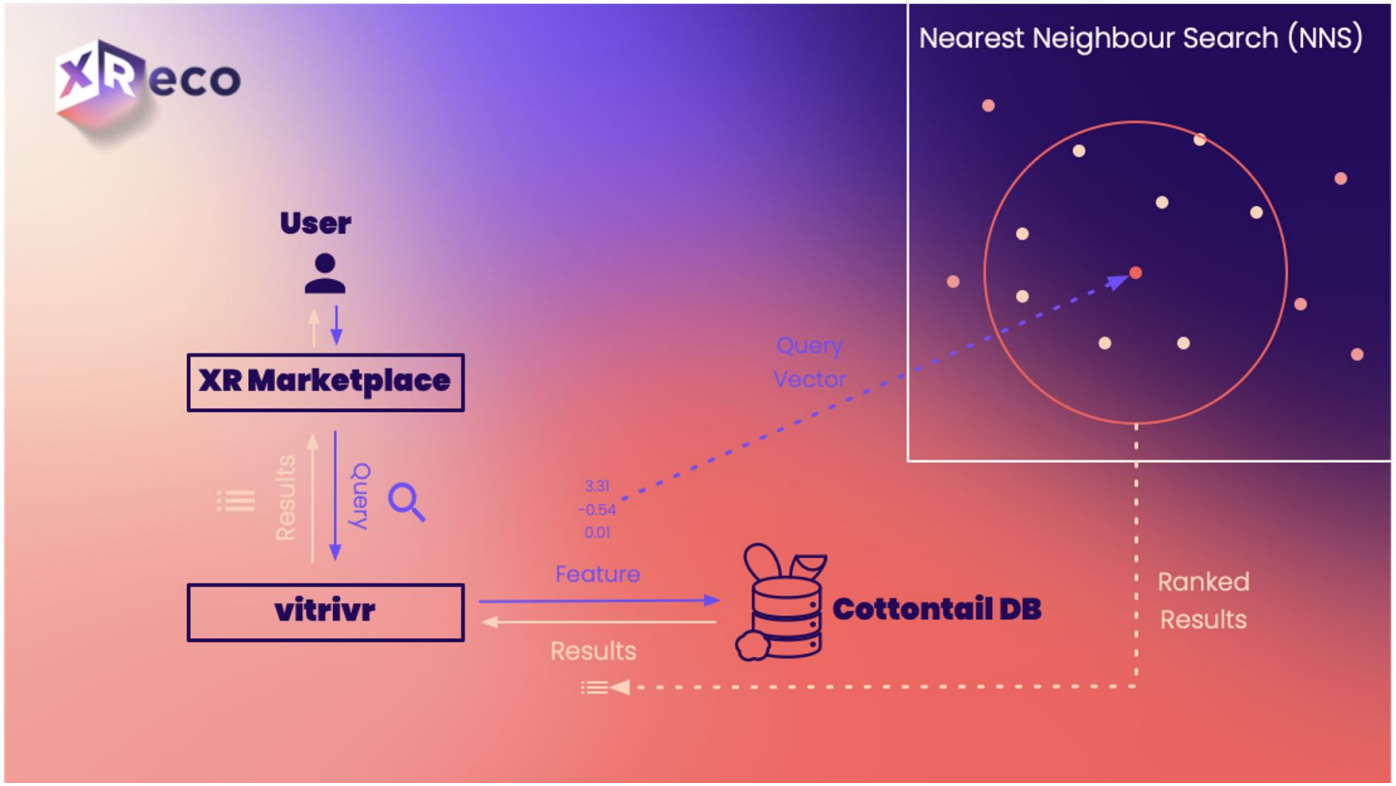


Image source: https://cdn.pixabay.com/photo/2023/03/20/15/40/cuba-7865319_960_720.jpg

Search in XR – Example XReco

Integration of vitrivr into XReco

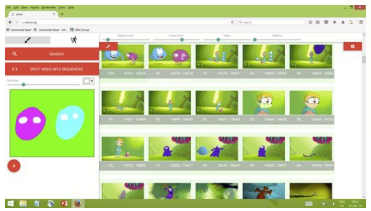


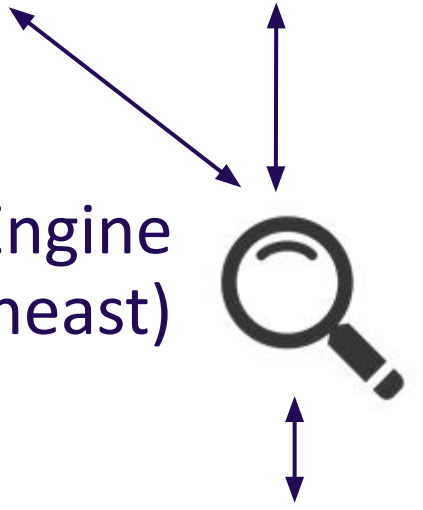
Search in VR – Example vitrivr-VR

VR front-end to vitrivr search system

- Query formulation
- Results presentation
- Relevance feedback

happens completely in an immersive way

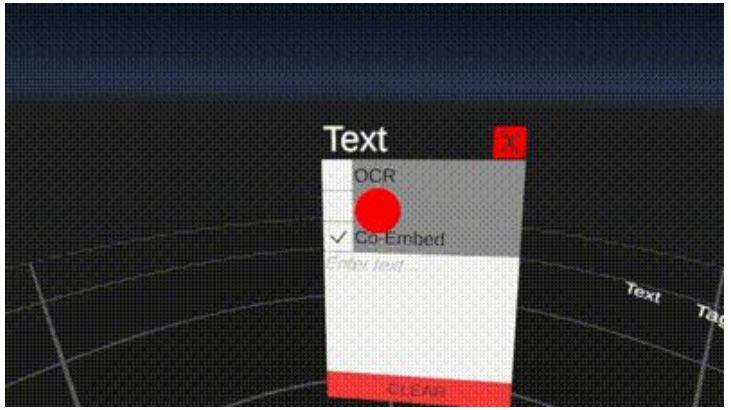


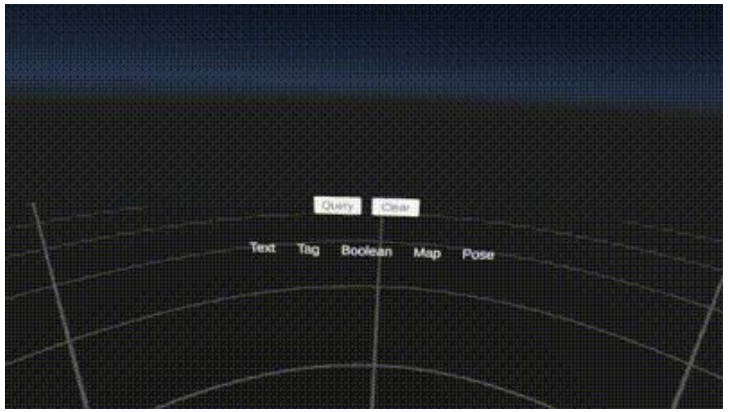


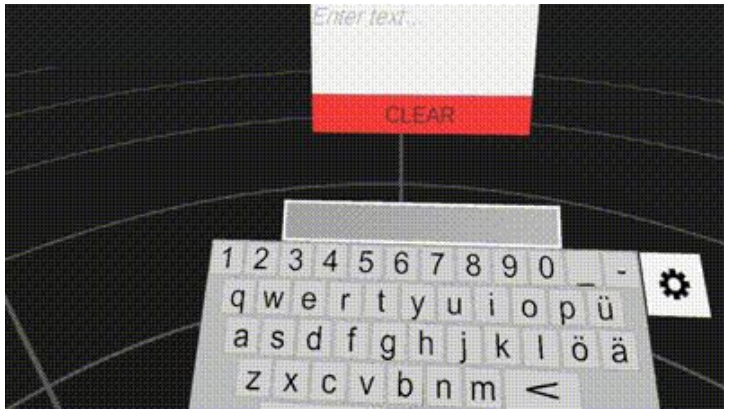
Search in VR: Query Formulation

Multimodal query formulation

- Text (Embedding, OCR)
- Boolean
- Geospatial
- Pose queries







Search in XR – Example GoFind!

- vitrivr front-end for smartphones
- Focus: cultural (architectural) heritage

Image sources: L. Sauter, Univ. Basel

Results Presentation

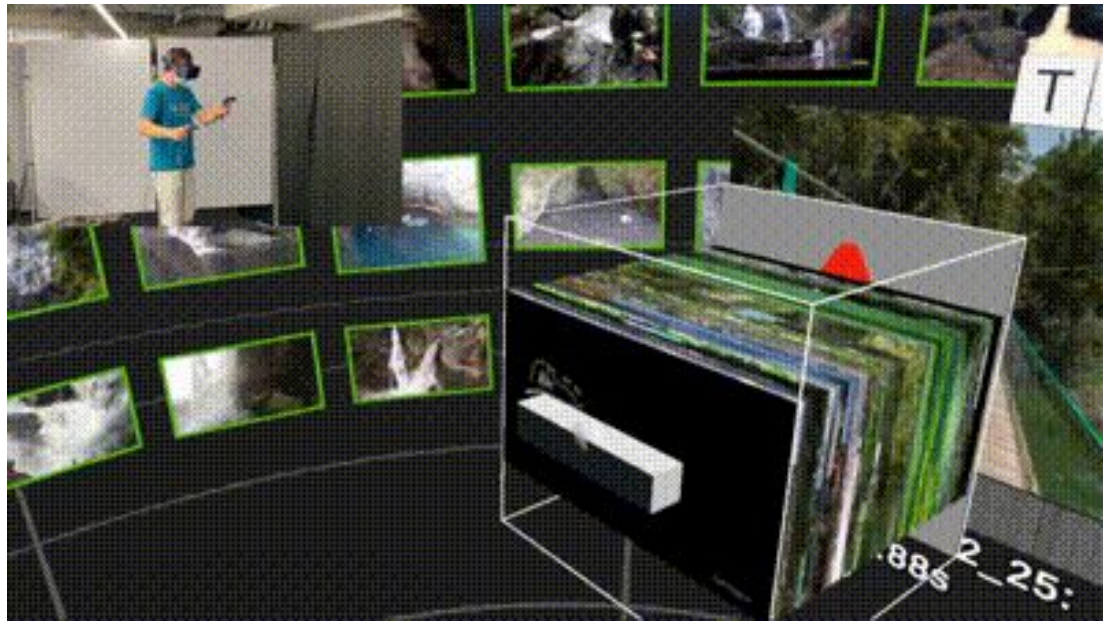
- Traditional (desktop) UIs suffer from a separation between user and content
- XR user interfaces allow for **immersive user experiences**
  - users are “within” a collection
  - integration of results into user’s view

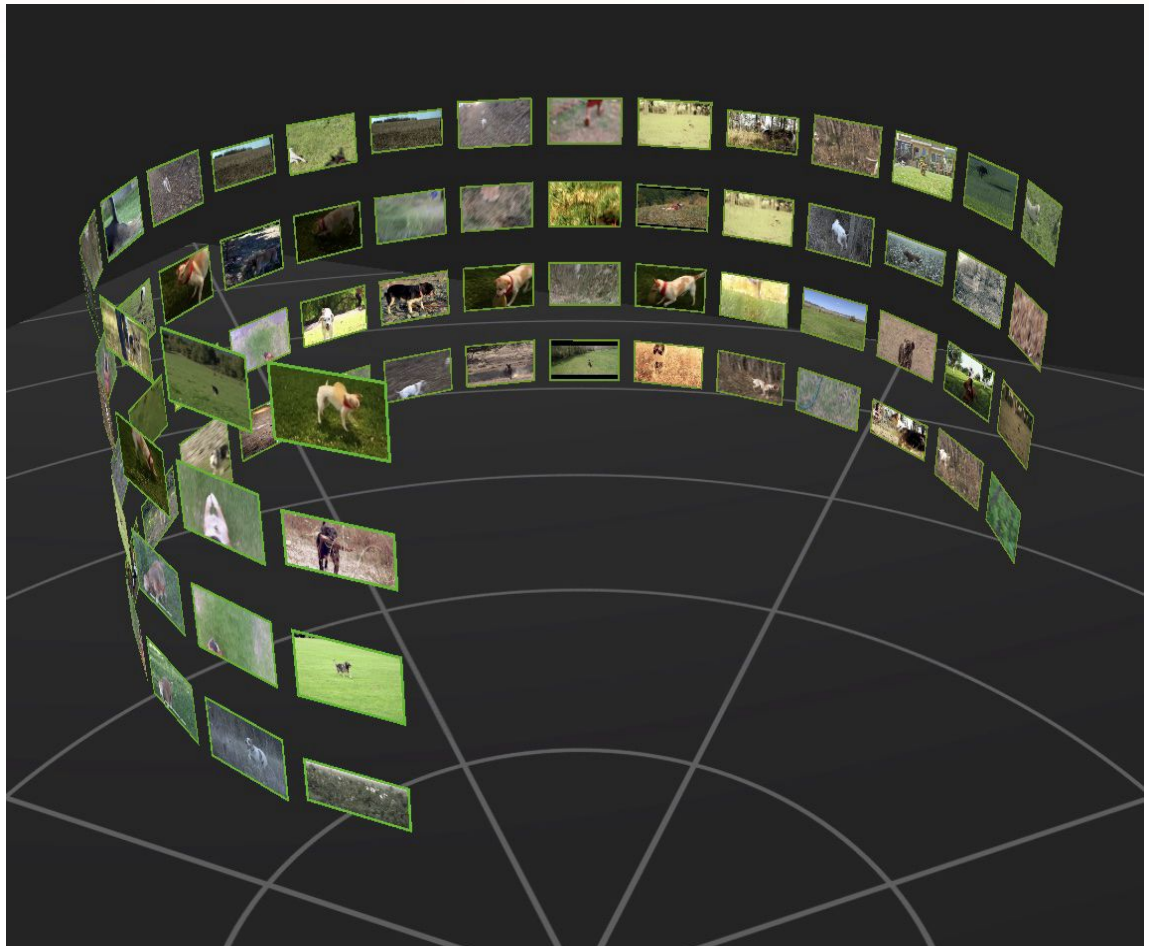
Image licensed from AdobeStock

Results Presentation in VR (vitrivr-VR) …

Cylindrical grid results display

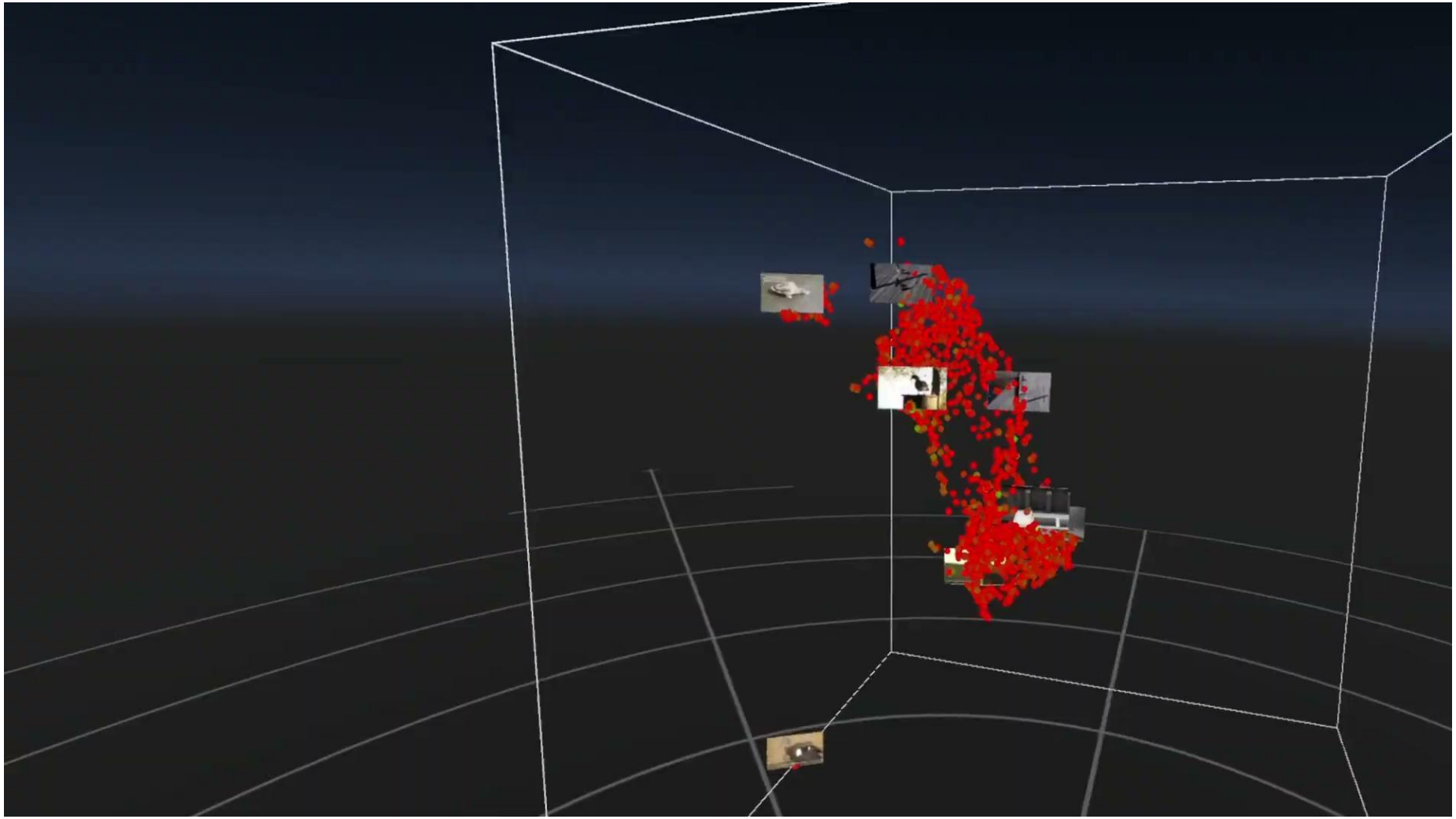
Multimedia drawer





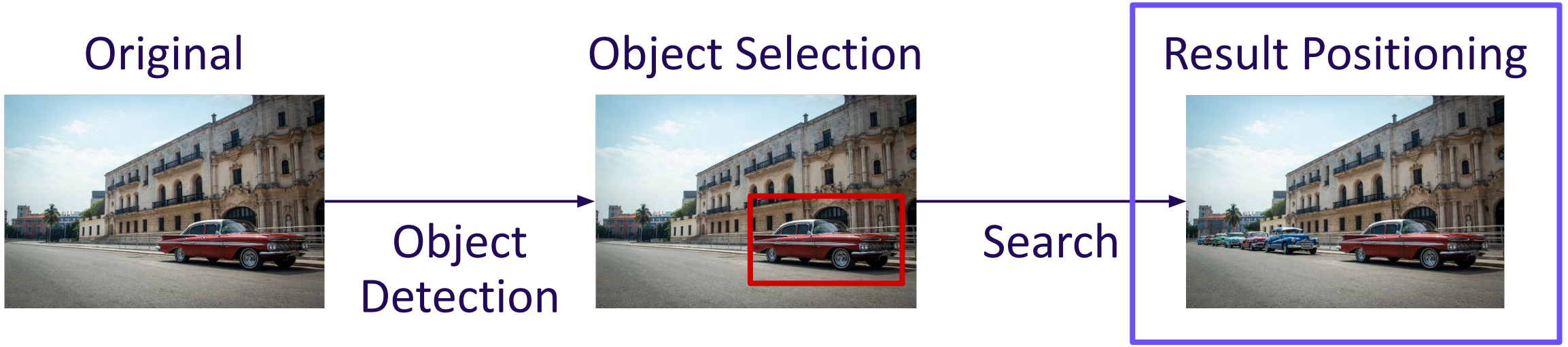
… Results Presentation in VR (vitrivr-VR)

Point cloud display from dimensionality reduction (collection analysis)



Results Presentation in XR

Integration of results in XR: semantically meaningful placement



Source: https://cdn.pixabay.com/photo/2023/03/20/15/40/cuba-7865319_960_720.jpg

2D Asset Retrieval for Reconstruction

Image from: Snavely et al. "Photo tourism: exploring photo collections in 3D." ACM SIGGRAPH 2006.

The hunger for 3D assets

- XR applications critically depend on good quality 3D representations of scenes and objects
- Marketplaces for objects exist
- Creating models for specific scenes or objects is a labour-intensive and costly process [Visu]
  - *A spaceship model: 30 – 32 hours*
  - *A rifle model: 110 hours*
  - *A sneaker model: 70 hours*
  - *An animation character: 110 hours to a couple of months*

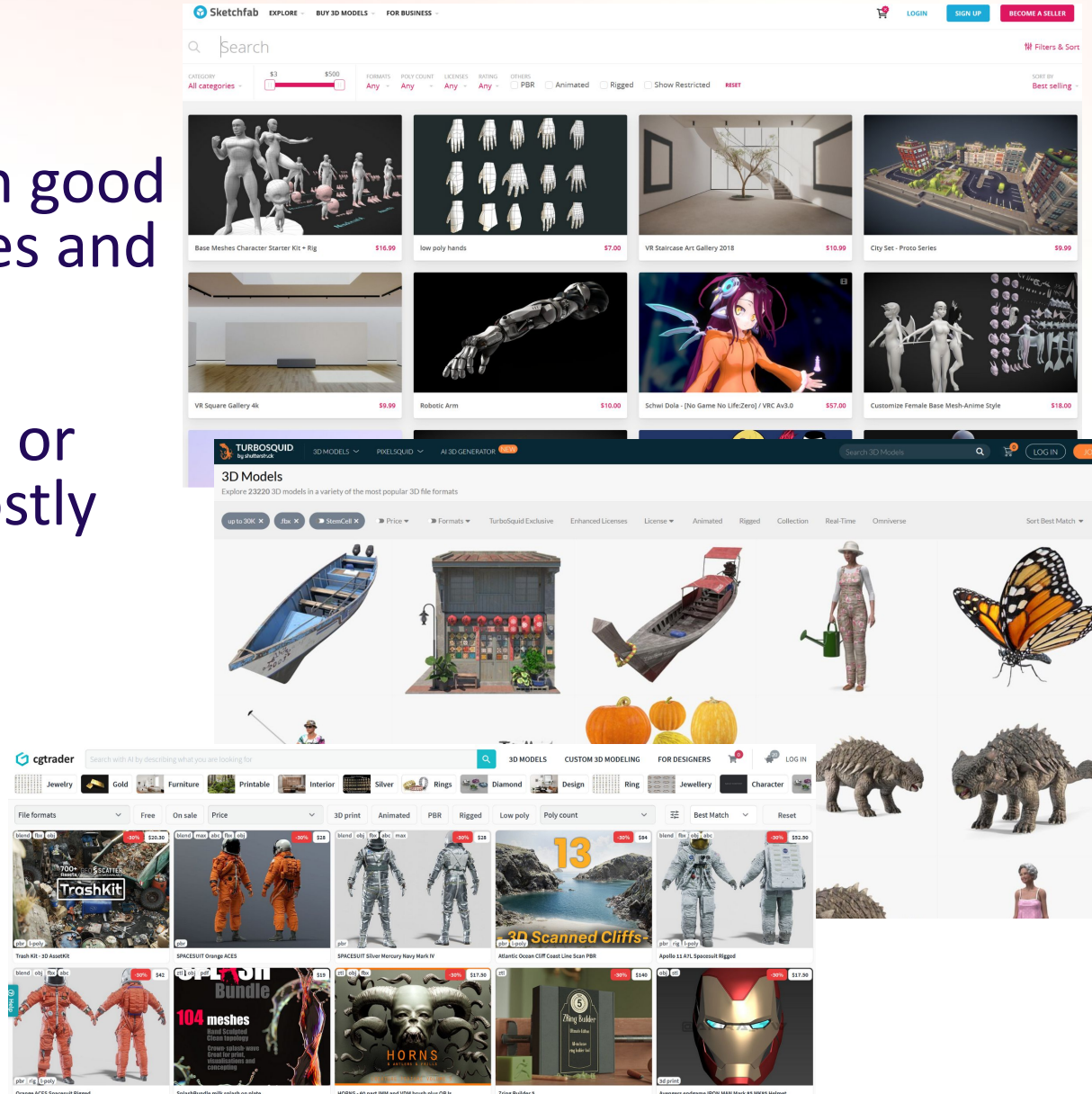


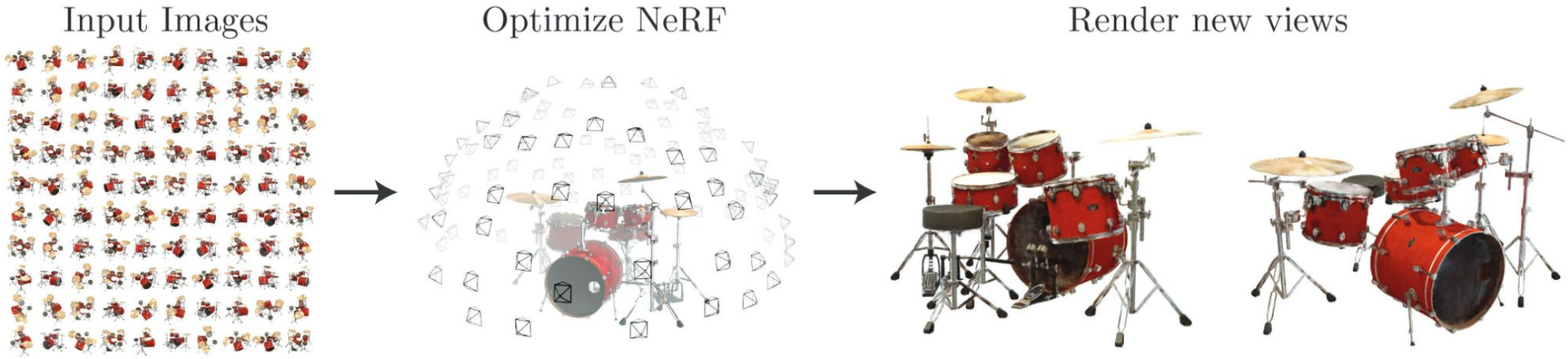
Image sources: sketchfab.com, turbosquid.com, cgtrader.com

From controlled to in the wild

- 3D reconstruction methods work well with material captured under controlled conditions
  - known poses/motion path
  - same condition of the object
  - same light conditions
  - same camera, lens and processing pipeline
- Content matching these conditions is scarce, or needs to be recorded specifically (takes time, costs money)
- But we have lots of “in the wild” content: image, broadcast and film archives, personal photo collections, …
- Challenges: **find the relevant content**, and **create good quality reconstructions**

Some recent advances in 3D reconstruction

- Neural Radiance Fields (NERFs) [M20]
  - relies on traditional methods such as SfM for pose estimation of input images (e.g. COLMAP)
  - overfit a neural network on a set of images to obtain radiance field
  - rendering novel views with good quality, but computationally costly
  - implicit representation, i.e. point cloud or mesh has to be derived



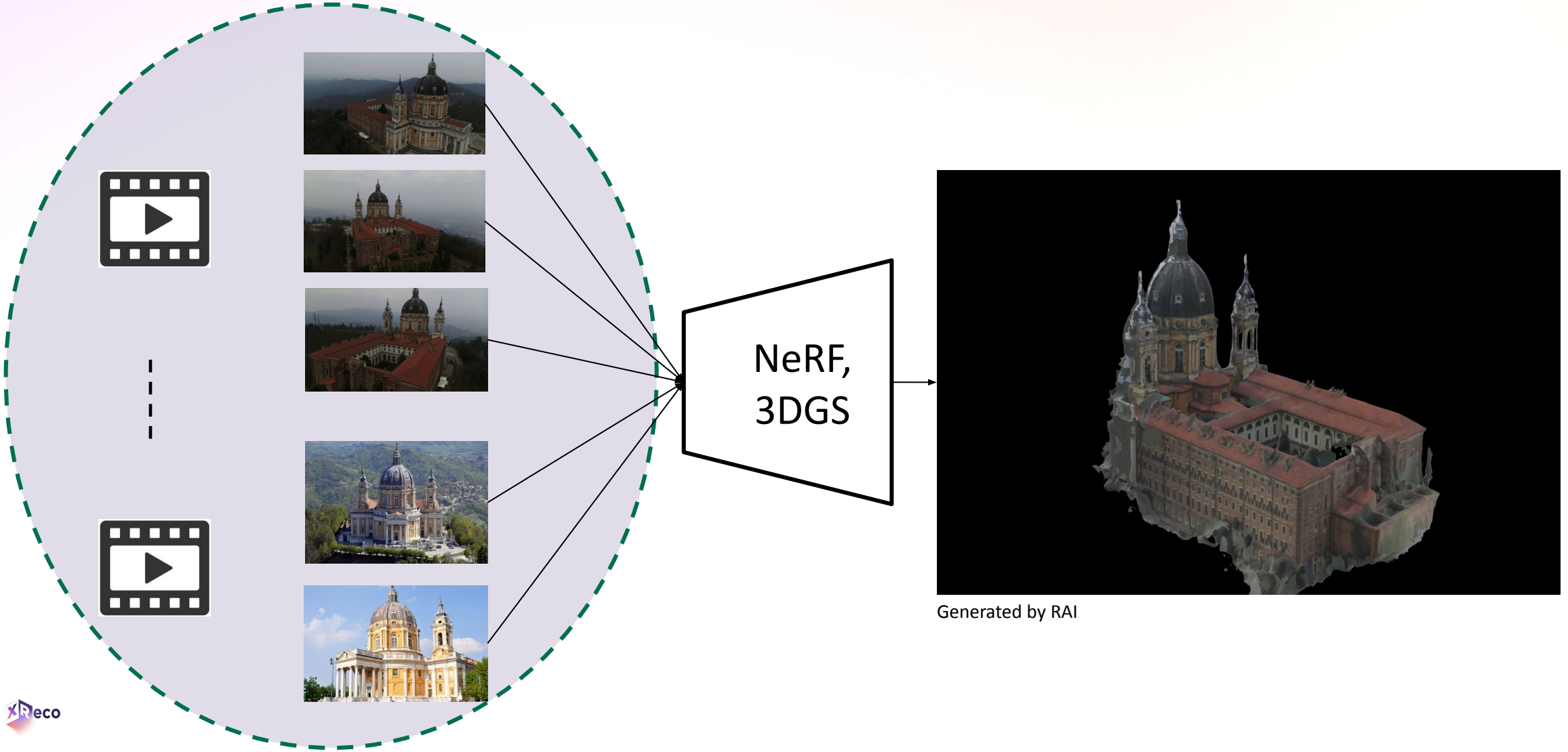
Some recent advances in 3D reconstruction

- 3D Gaussian Splatting [K23]
  - create point cloud using traditional method such as SfM
  - learn parameters of a 3D Gaussian around each point of the point cloud

Image from [K23]

Image from https://huggingface.co/blog/gaussian-splatting

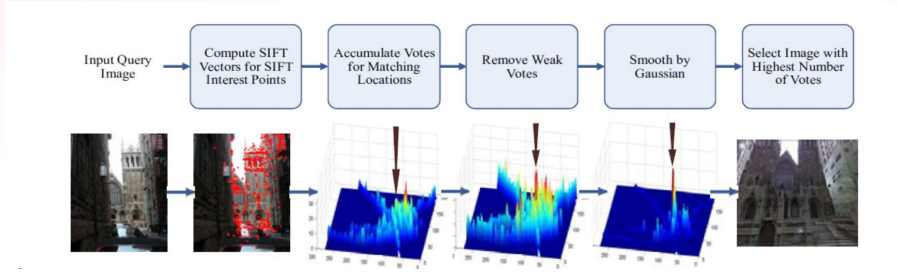
Vision

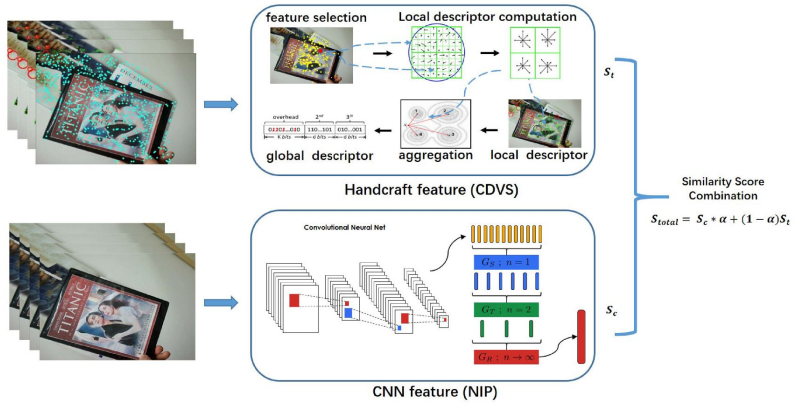


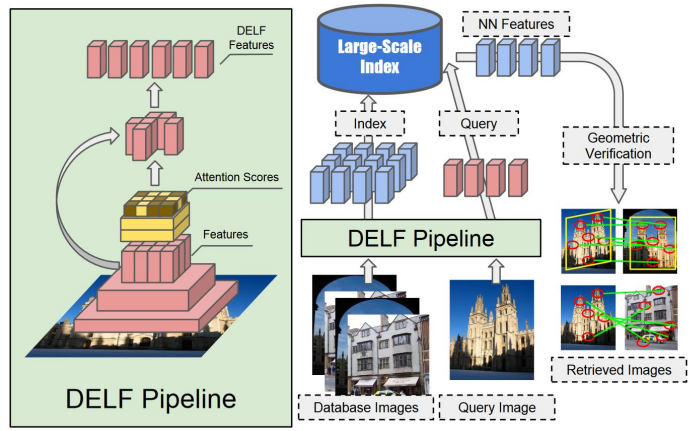
Generated by RAI

Mining landmark images: approaches

- Traditional hand crafted features (SIFT, SURF, MSER, etc.) [Z10, C11]
- Moving local feature approach to deep learning, e.g. DELF (DEep Local Features) [N17]
- Hybrid approaches
  - fusing traditional and learned features
  - e.g. MPEG CDVA [D18]

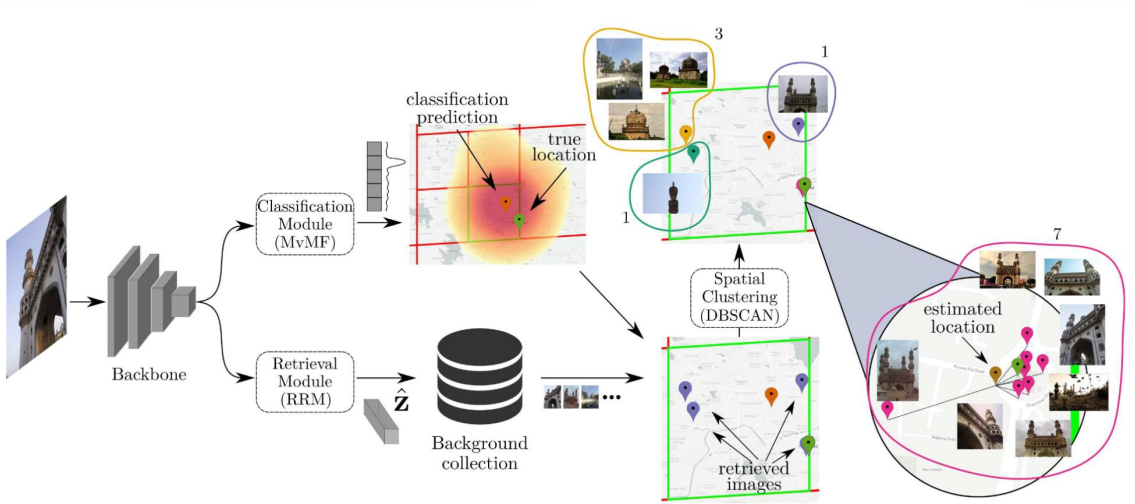


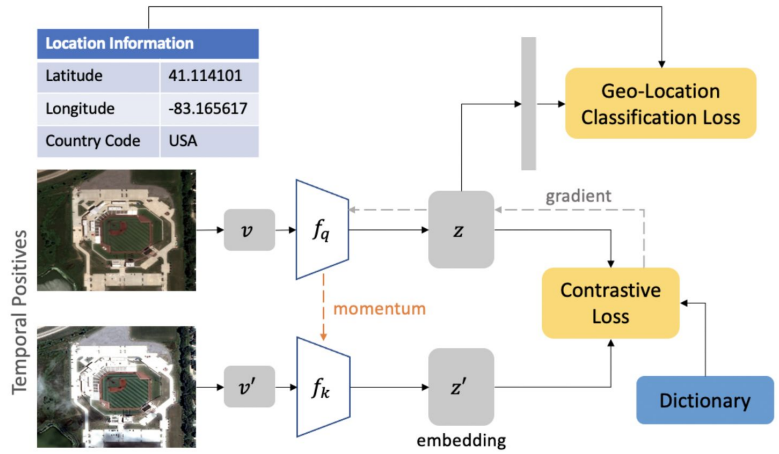




Mining landmark images: approaches

- CNN-based approaches
  - lightweight landmark recognition model using EfficientNet and Linear Discriminant Analysis (LDA) [R23]
  - ResNets with increasing image resolution [Y22]
- Contrastive and self-supervised learning
  - EfficientNet and contrastive learning for classification on a grid [K21]
  - for satellite images [A21]





Landmark datasets

- Early datasets: Oxford [P07] and Paris [P08]
- Increasing size and diversity: Pittsburg250k [T13], Tokyo 24/7 [T15]
- Google Landmarks v1 [N17] and v2 [W20a]
- Datasets from autonomous driving research context BDD100k [Y18], Mapillary street level sequences [W20b]

Benchmarks

- MediaEval [ME] organised a placing task from 2010-2016
- Google Landmark Recognition Retrieval Challenges
  - 2018-2021 [LR18, LR19, LR20, LR21]
  - moved to instance-level recognition [ILR], including language-based embeddings
- Deep Visual Geo-localization Benchmark [B22]

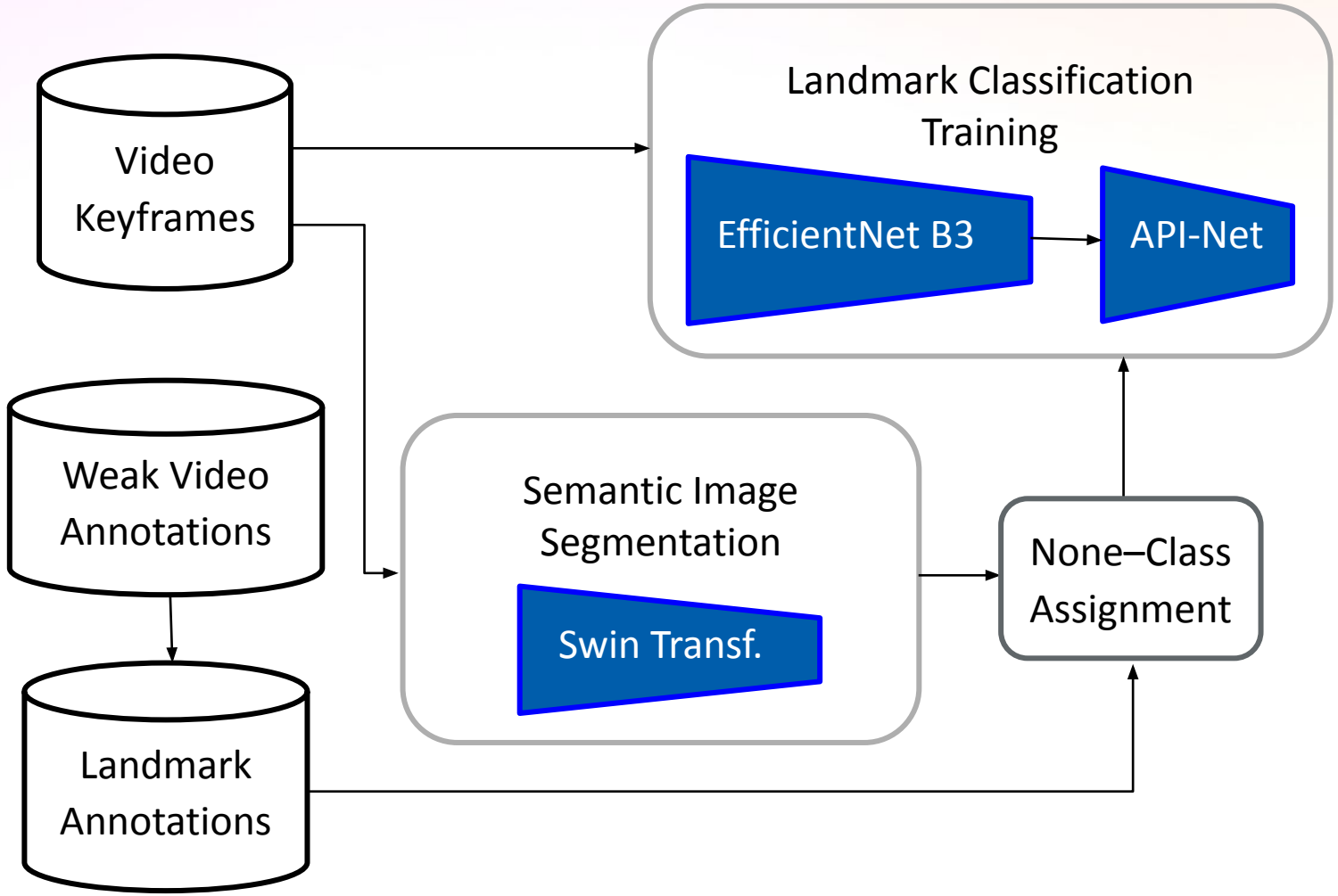
Issues with landmark retrieval methods

- Benchmarks drive progress, methods align with their setup
- Existing landmark recognition benchmarks
  - are based on images
  - assume correct labelling of the training samples
- Available data
  - videos are a very useful source for reconstruction, as they contain different views taken under the same conditions
  - they are typically not annotated on a shot (or even frame) granularity
  - relevant landmark (or even only city) may be annotated on the level of an entire programme (at best, on a story)

Approach to leverage archive content

- Using weakly annotated videos to gain training data for recognition of landmarks
  - Framework for selecting training data
    - from videos by semantic image segmentation and
    - by using images mined from web searches
- Train a fine-grained landmark classifier from a small set of training samples per class
- Enable adding landmarks incrementally in an efficient way
  - address requirements coming up while working on an XR production

Swin+API-Net

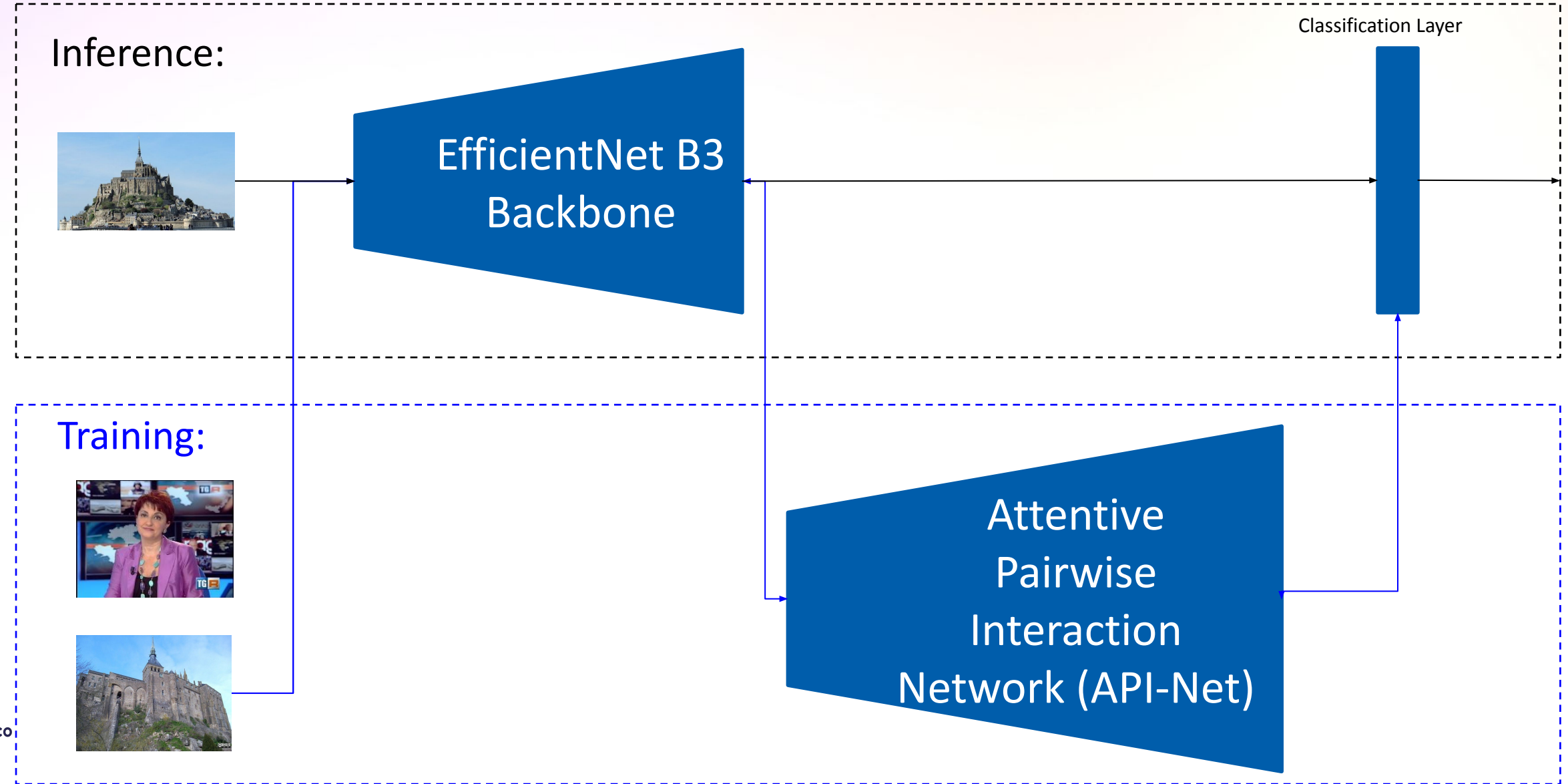


API-Net: Attentive Pairwise Interaction Network
Swin Transformer: Hierarchical Vision Transformer using Shifted Windows

None class example

Landmark example

Fine grained Image Classification with Attentive Pairwise Interaction Network (API)



Training Data Generation by Web Image Mining (CDVA+API-Net)

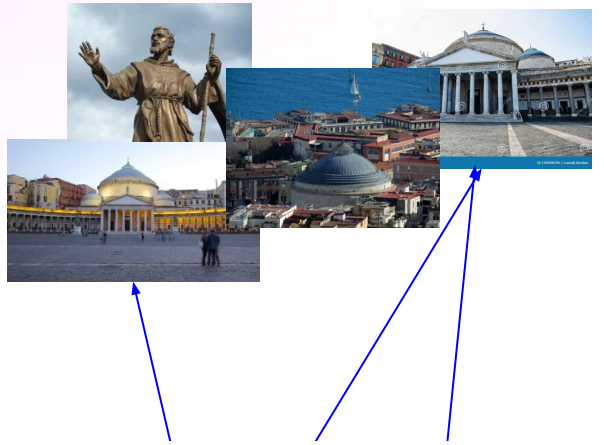
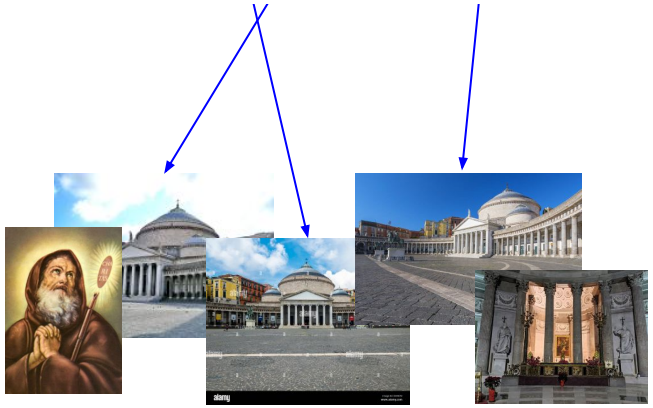
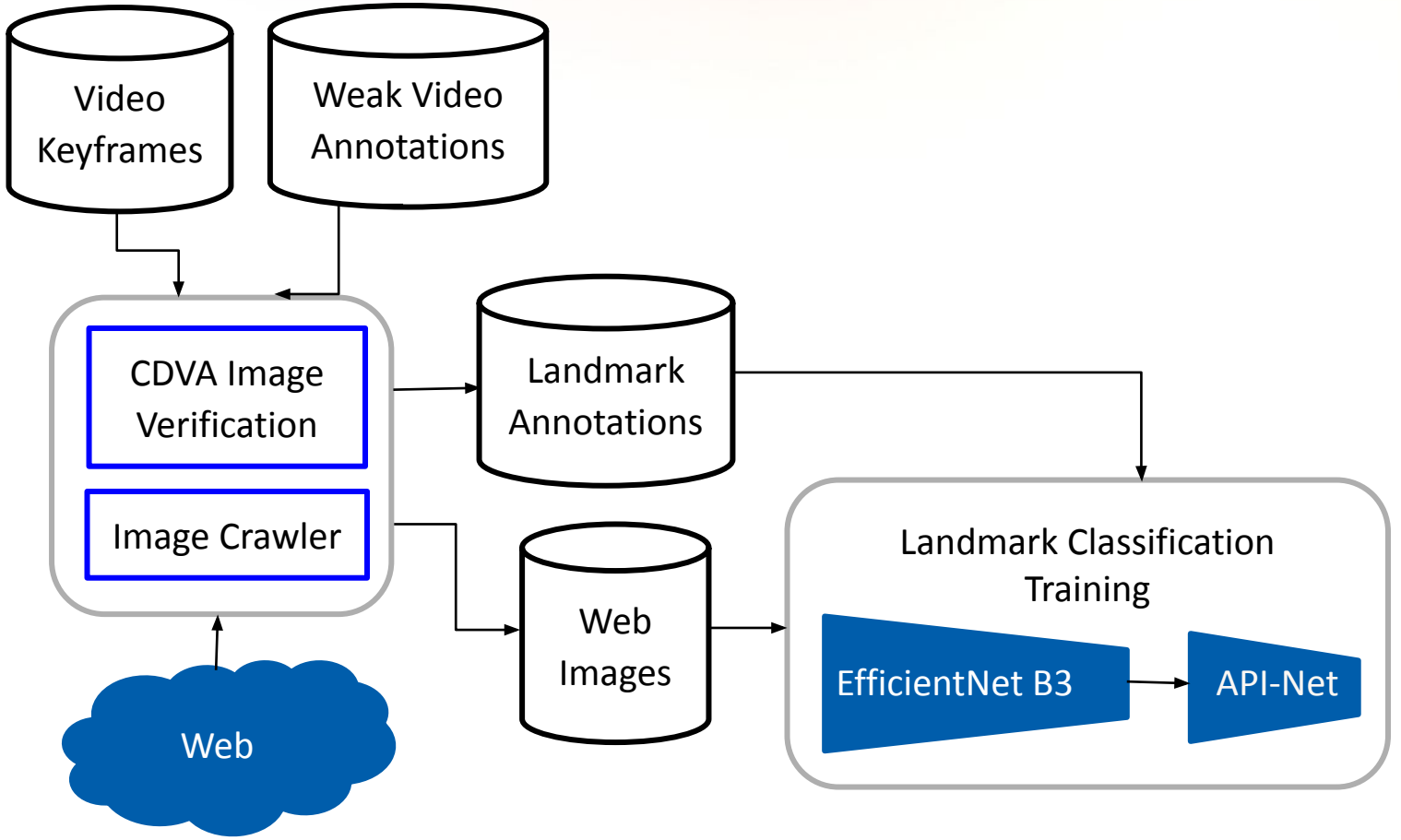


Image similarity





CDVA: MPEG Compact Descriptors for Video Analysis

Weakly Annotated Video Landmark (WAVL) dataset

- WAVL dataset
  - Merged from two different sources:
    - Google Landmarks v2 image dataset
    - Vimeo Creative Commons Collection (V3C) video dataset
    - https://github.com/XRecoEU/WAVL-Dataset
  - Similar characteristics as the RAI dataset (number of landmarks, noise images, …)

Dataset	LM	LM Images	Noise Images	Mean LM Images	Std. Dev. LM Images	Mean Noise Images	Std. Dev. Noise Images
RAI	163	5620	1871	34.55	26.92	10.63	10.60
WAVL	141	4230	1592	29.42	1.33	11.29	1.07

LM: landmarks

Evaluation with the WAVL Dataset

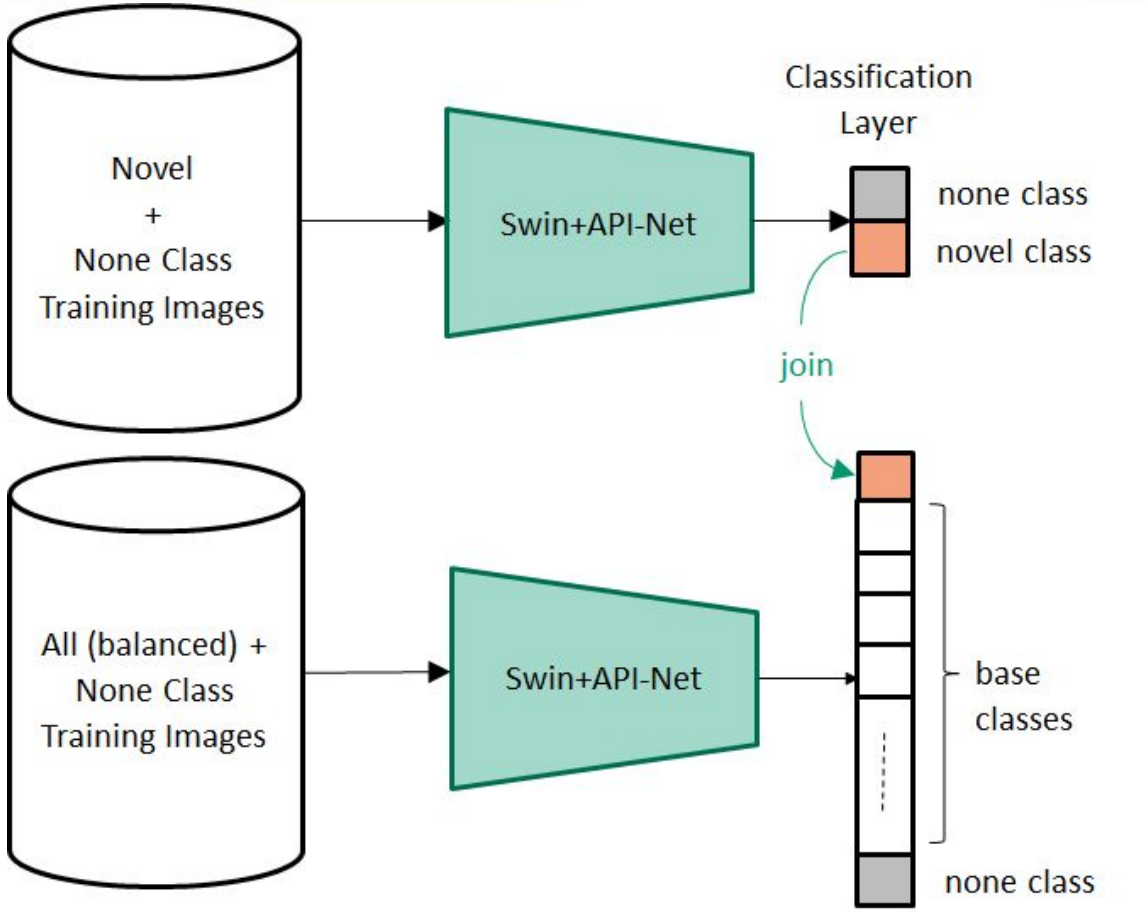
- Baseline method 1: smlyaka
  - First place in the Google Landmark Retrieval 2019 Challenge
  - ResNet-101 and spatial verification with Deep Local Attentive Features (DELF)
  - filtering method proposed in this approach turned out to reduce performance
- Baseline method 2: API-Net without pre-filtering images

Method	GAP	Threshold	Precision	Recall	SBA
smlyaka	47.34	2.0	53.56	**66.31**	54.64
API-Net	37.43	9.0	56.27	32.22	53.99
Swin+API-Net	**53.86**	6.5	69.63	62.06	68.96
CDVA+API-Net	53.03	6.0	**79.77**	52.57	**72.43**

GAP: global average precision SBA: symmetric balanced accuracy

Class incremental learning

- Starting from an initial network trained with the base classes
- Training of the classification layer for novel classes
- Join networks by adding nodes of the new classes to the initial classification layer
- Fine-tuning with balanced training set from all classes
  - following the two-stage fine-tuning approach [W20]



Class incremental learning

- Classification performance similar to full set
  - results on RAI Monuments of Italy dataset [N24b]

Number base landmarks	Number novel landmarks	MAP all	GAP all	Threshold	BA all	SBA all
162	1	57.82	50.94	7.0	64.38	64.15
158	5	55.56	52.57	7.0	65.79	65.57
153	10	52.99	49.88	7.0	63.59	63.92
143	20	53.91	50.68	7.0	61.24	61.54

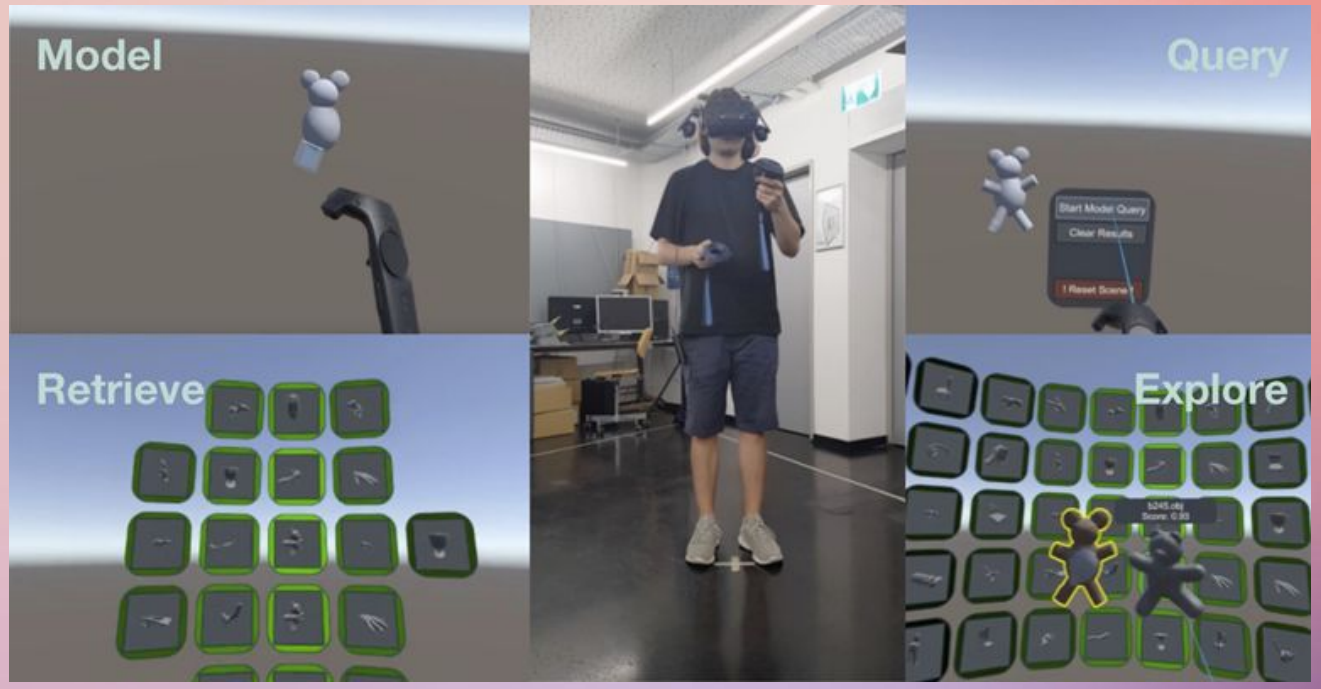
- Fast and interactive training is possible
  - when training the classification head only, ~100s on a recent consumer GPU

3D assets from 2D archive content

- approaches to mine relevant images from large scale content sets exist
- retrieval results still often return very heterogeneous content sets
- focus on some “postcard views” of the landmark rather than covering all views
- obtaining high quality 3D assets remains still challenging

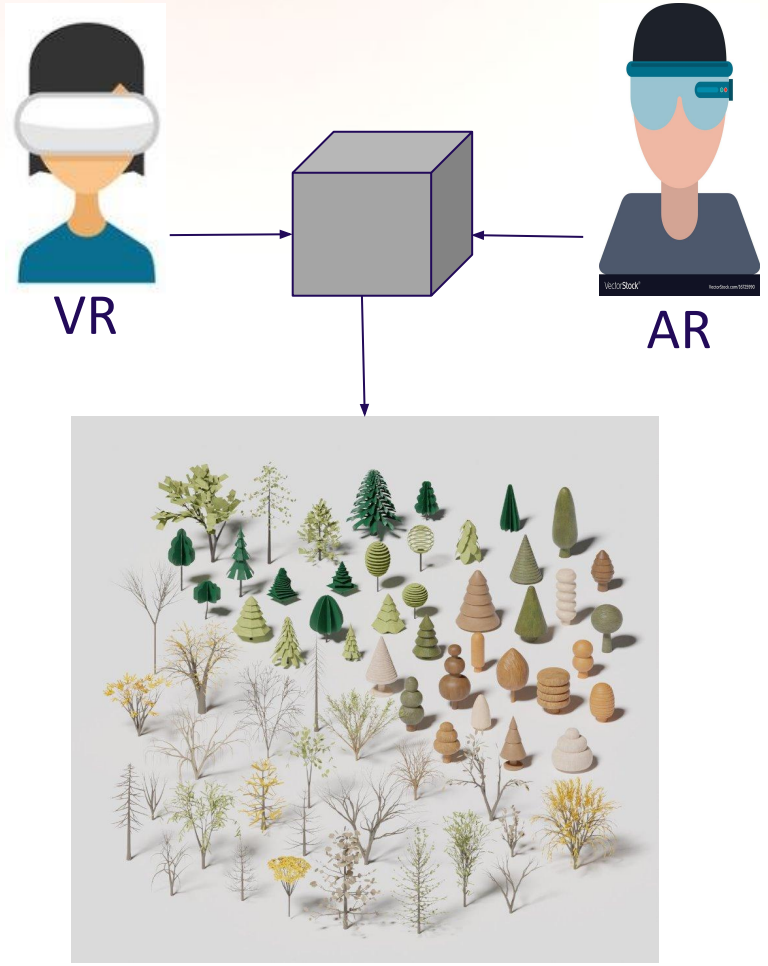
3D Object Retrieval

Importance, Applications & Methodologies



The Power of 3D Object Retrieval (1)

- XR technologies (e.g., VR, AR)
  - Interaction with digital 3D data



https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.vectorstock.com%2Froyalty-free-vector%2Fsmart-glasses-icon-image-vector-16725990&psig=AOvVaw03fLP__sDxP3rKiT34jmb8&ust=1715847264048000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCJC40M2bj4YDFQAAAAAdAAAAABBO
https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.123rf.com%2Fclipart-vector%2Fvr_headset_group.html&psig=AOvVaw1FIzFztkgCkMgL-FVB1Gm8&ust=1715847237392000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCOCiopubj4YDFQAAAAAdAAAAABAy
https://www.google.com/url?sa=i&url=https%3A%2F%2Ftoffu.co%2Fproducts%2F3d-model-trees-bundle&psig=AOvVaw0jONsSdOsBmk0srzTRHQtz&ust=1715865092514000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCMjai-zij4YDFQAAAAAdAAAAABAE

The Power of 3D Object Retrieval (2)

- Importance in XR experiences
  - Virtual World closer to Real World

https://www.google.com/url?sa=i&url=https%3A%2F%2Fdocs.unity3d.com%2F2018.1%2FDocumentation%2FManual%2Fterrain-Trees.html&psig=AOvVaw29pLYtQ7aiWXplZlIEiESj&ust=1715869624784000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqGAoTCKDkiM7uj4YDFQAAAAAdAAAAABDgAQ

Applications

- VR Gaming
- AR Shopping
- Educational Simulations
- Medical Training
- Tourism & Travel Experiences
- etc.

https://designer.microsoft.com/image-creator

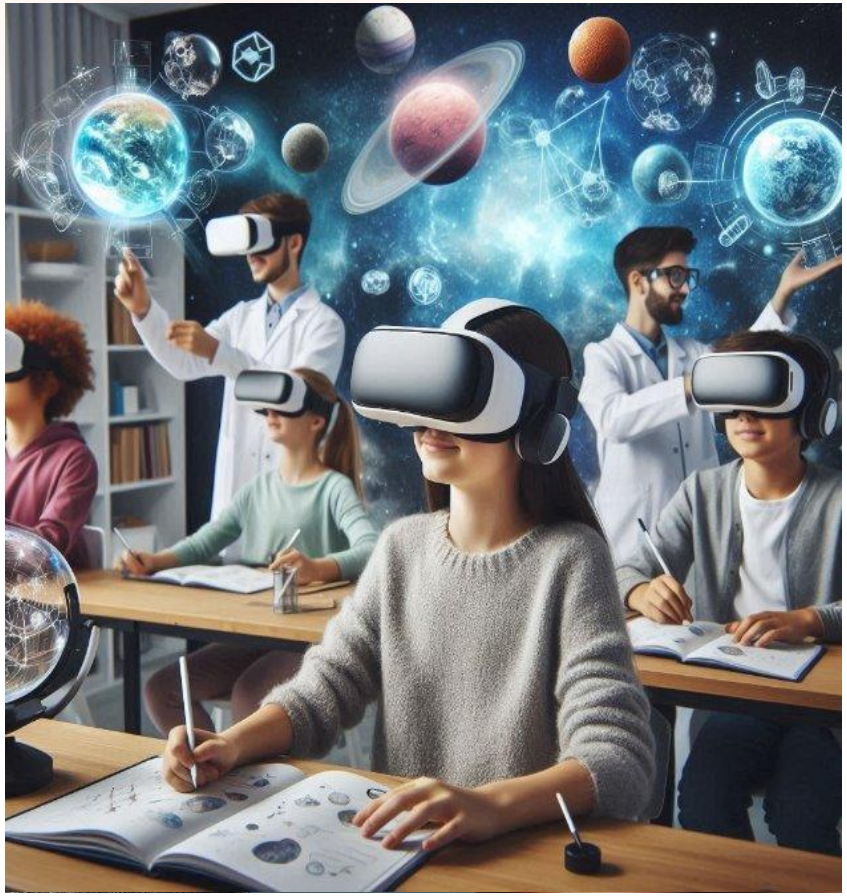
Applications

- VR Gaming
- AR Shopping
- Educational Simulations
- Medical Training
- Tourism & Travel Experiences
- etc.

https://designer.microsoft.com/image-creator

Applications

- VR Gaming
- AR Shopping
- Educational Simulations
- Medical Training
- Tourism & Travel Experiences
- etc.



https://designer.microsoft.com/image-creator

Applications

- VR Gaming
- AR Shopping
- Educational Simulations
- Medical Training
- Tourism & Travel Experiences
- etc.

https://www.google.com/url?sa=i&url=https%3A%2F%2Fsmarttek.solutions%2Fblog%2Fvr-for-surgery%2F&psig=AOvVaw1yTTAFTIuuceAorKWkdVZr&ust=1715929663733000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCKjptKLOkYYDFQAAAAAdAAAAABAE

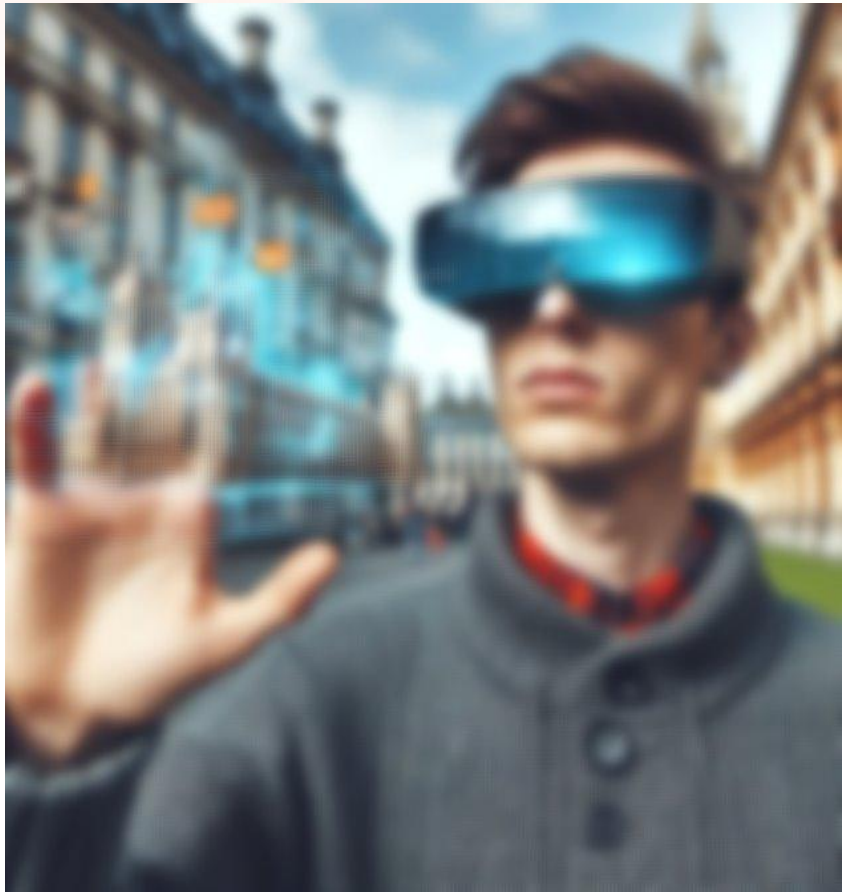
Applications

- VR Gaming
- AR Shopping
- Educational Simulations
- Medical Training
- Tourism & Travel Experiences
- etc.

https://designer.microsoft.com/image-creator

Applications

- VR Gaming
- AR Shopping
- Educational Simulations
- Medical Training
- Tourism & Travel Experiences
- etc.



https://designer.microsoft.com/image-creator

Search & Retrieve 3D data

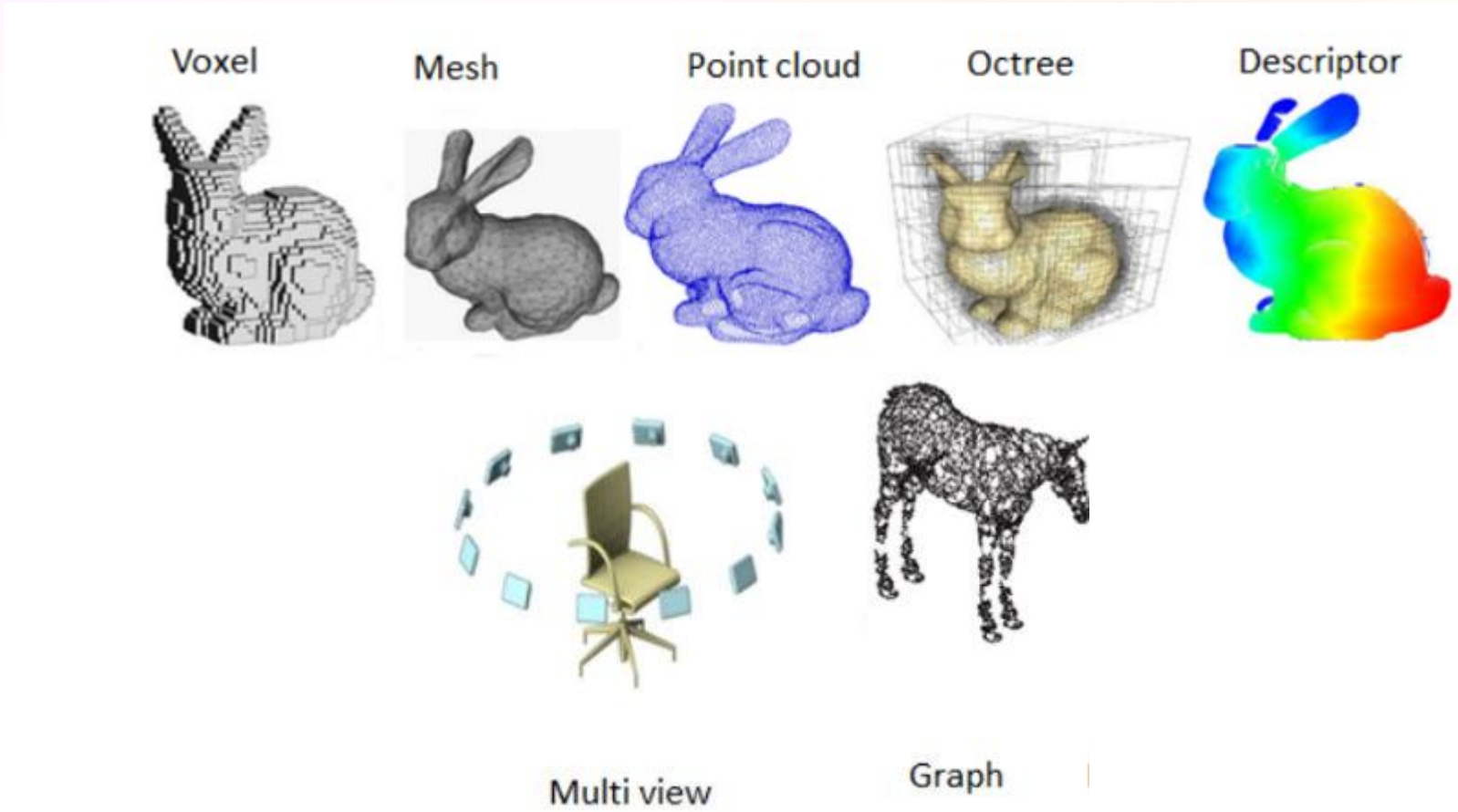
Query

DB

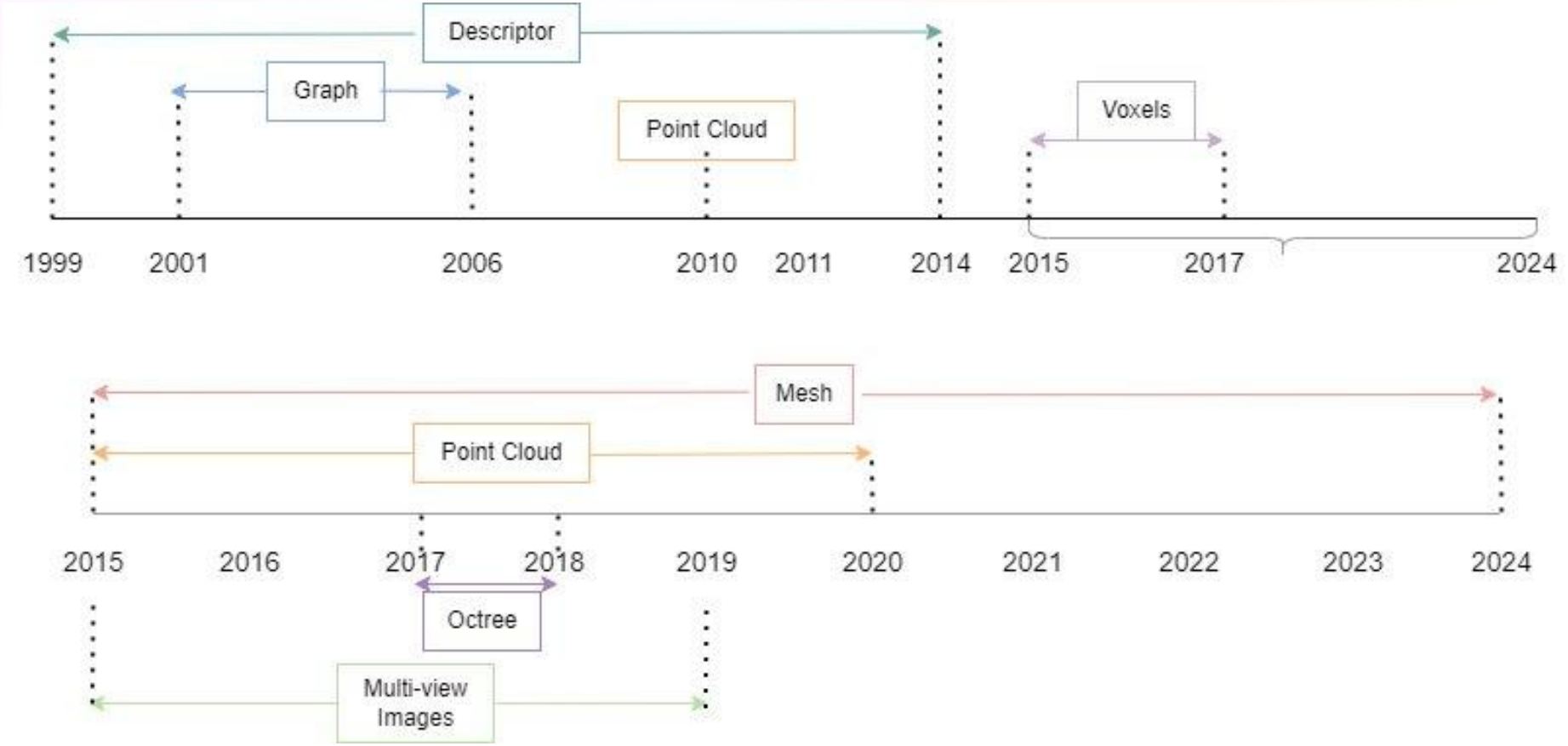
Retrieved list

https://www.google.com/url?sa=i&url=https%3A%2F%2F3dmodels.org%2F3d-models%2Foak-tree%2F&psig=AOvVaw0jONsSdOsBmk0srzTRHQtz&ust=1715865092514000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCMjai-zij4YDFQAAAAAdAAAAABAQ
https://www.google.com/url?sa=i&url=https%3A%2F%2Fwww.itapetinga.ba.gov.br%2F%3Fs%3Doak-tree-3d-model-cc-dNzKGXUP&psig=AOvVaw0jONsSdOsBmk0srzTRHQtz&ust=1715865092514000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCMjai-zij4YDFQAAAAAdAAAAABBB
https://www.google.com/url?sa=i&url=https%3A%2F%2F3dmodels.org%2F3d-models%2Fenglish-oak-tree%2F&psig=AOvVaw0jONsSdOsBmk0srzTRHQtz&ust=1715865092514000&source=images&cd=vfe&opi=89978449&ved=0CBIQjRxqFwoTCMjai-zij4YDFQAAAAAdAAAAABBL

Representation/Modalities (1) [Gezawa20]

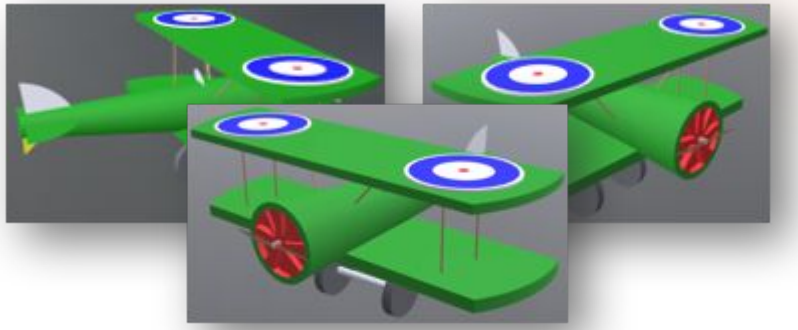


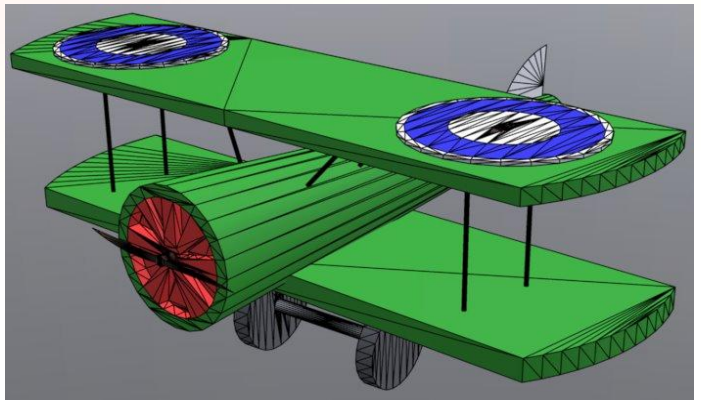
Representation/Modalities (2) [Gezawa20]

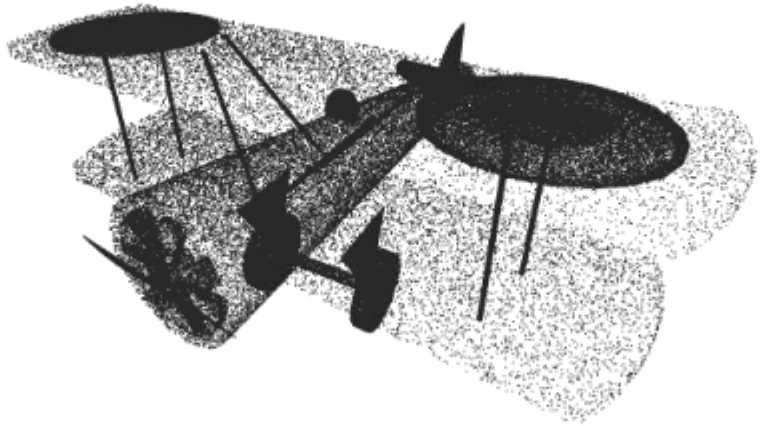


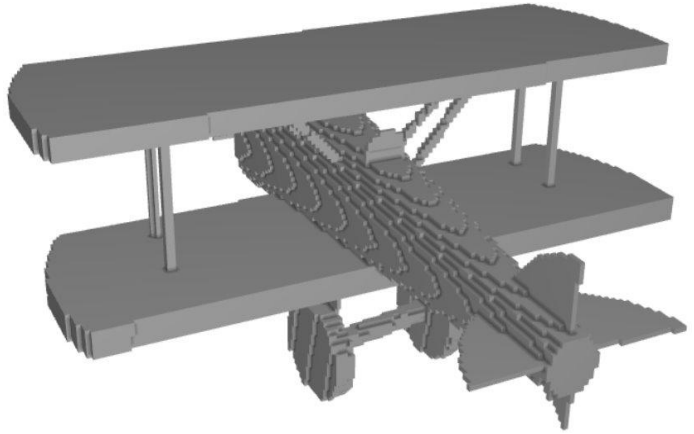
Representation/Modalities (3)

Type
Multi-view images **(visual)**
Meshes **(spatial)**
Point Clouds **(spatial)**
Voxels **(spatial)**









Representation/Modalities (4)

Type	Pros	Cons
Multi-view images (visual)	• Rich Visual Information • Compatibility	• Data Acquisition • Storage • Sensitivity to Occlusion
Meshes (spatial)	• Geometric Accuracy • Flexibility	• Complexity • Artifacts
Point Clouds (spatial)	• Precise Geometry • Scalability	• Density Variation • Limited Surface Information
Voxels (spatial)	• Structured and Uniform • Simulation Compatibility	• High Memory Consumption • Scalability Issues • Complex Smooth Surfaces

Challenges

- Computational and Memory Complexity
- Insufficient Invariance to Transformations
- Multimodal Integration Complexity
- Limited labeled 3D data

Challenges

- Computational and Memory Complexity
- Insufficient Invariance to Transformations
- Multimodal Integration Complexity
- Limited labeled 3D data

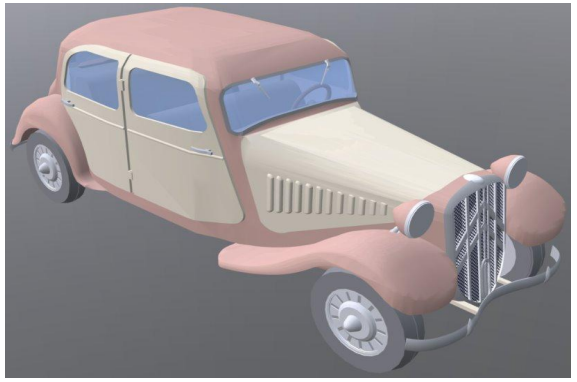
Idea

Metaverse Apartment Retrieval Challenge
(CV4Metaverse workshop, ECCV 2024,Milano, Italy)
Ask: Alex Falcon, Ali Abdari

Scan me!

Benchmark Datasets

Type	Size	Labels	Modalities		
			Image	Mesh	Point Cloud
ModelNet10	4,899	10	✓	✓	✓
ModelNet40	12,311	40	✓	✓	✓
ShapeNetCore	31,854	48	✓	✓	✓
BuildingNet_v0	2,000	60	✓	✓	✓
XRECO.Buildings.Monuments	201	12	✓	✓	✗



https://modelnet.cs.princeton.edu/#
https://huggingface.co/datasets/ShapeNet/ShapeNetCore
https://buildingnet.org/
https://zenodo.org/records/10809451

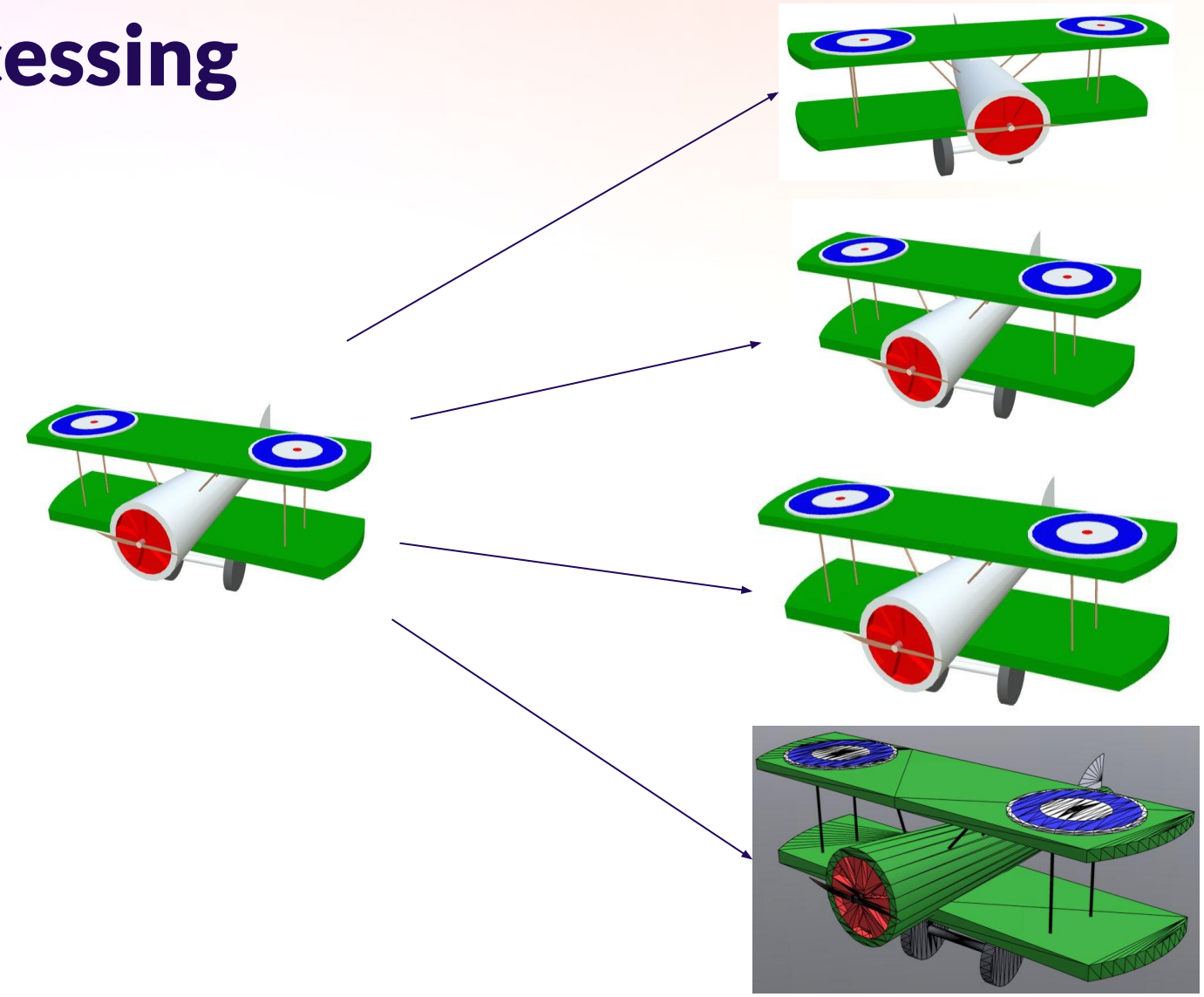
Sketchfab 3D Creative Commons Collection (S3D3C)

- Most existing 3D collections focus on model shapes
- S3D3C aims at providing advanced materials, textures, and animations

	CC0		CC-BY		CC-BY-SA		CC-BY-ND		Total	
Models	3,858		14,819		13,457		8,668		40,802	
Texture file	3,150	81.6%	13,080	88.3%	9,927	73.8%	5,176	59.7%	31,333	76.8%
Animation	67	1.7%	1,612	10.9%	441	3.3%	417	4.8%	2,537	6.2%
Sound	10	0.3%	279	1.9%	81	0.6%	9	0.1%	379	0.9%
Has description	3,712	96.2%	13,428	90.6%	9,790	72.8%	4,794	55.3%	31,724	77.8%
Has tag	3,440	89.2%	14,139	95.4%	12,152	90.3%	7,081	81.7%	36,812	90.2%
Has category	3,570	92.5%	13,602	91.8%	8,651	64.3%	4,909	56.6%	30,732	75.3%
Age restricted	0	0.0%	98	0.7%	0	0.0%	0	0.0%	98	0.2%

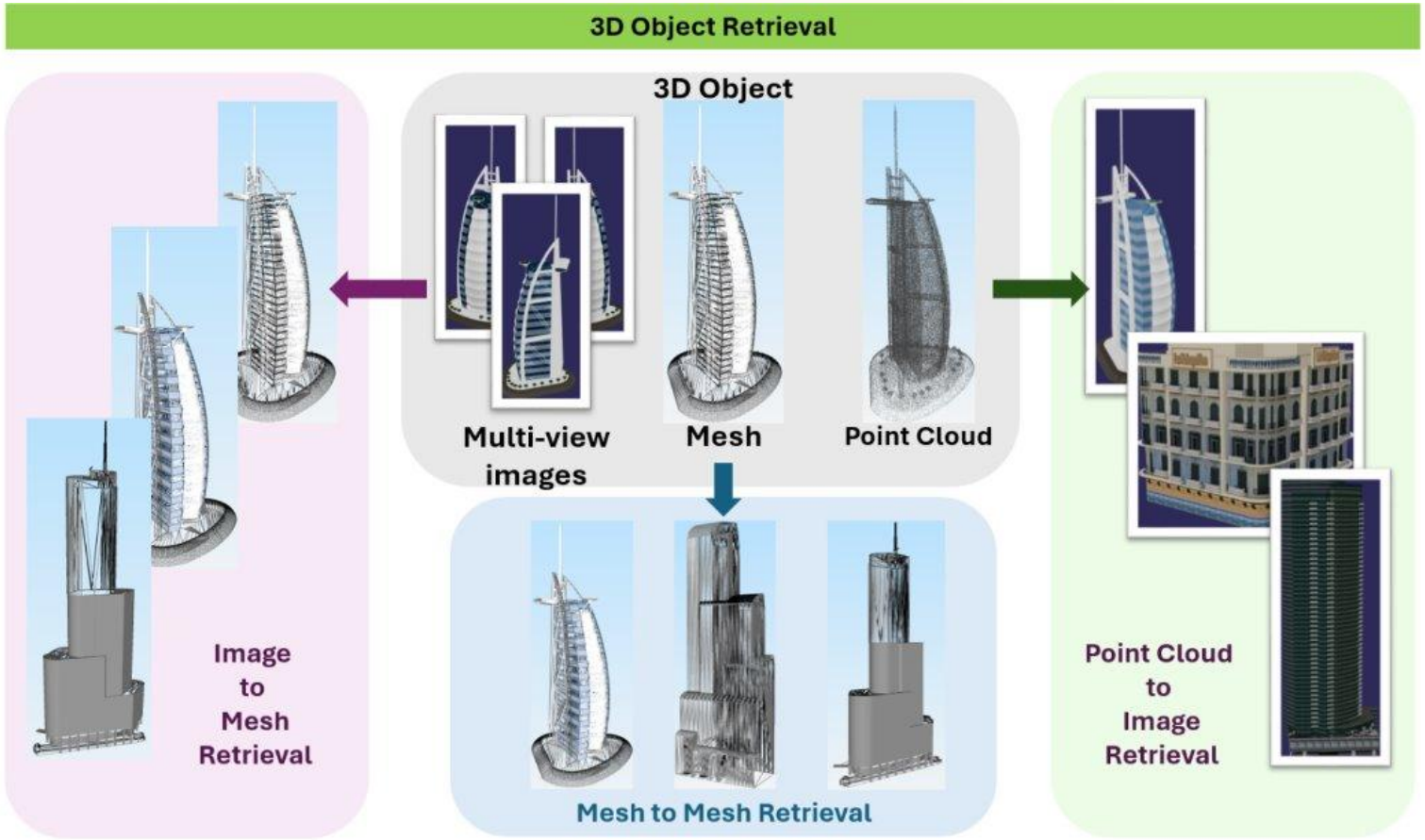
Data Preprocessing

- Rotation
- Translation
- Scaling
- Simplification



SOTA Approaches (1)

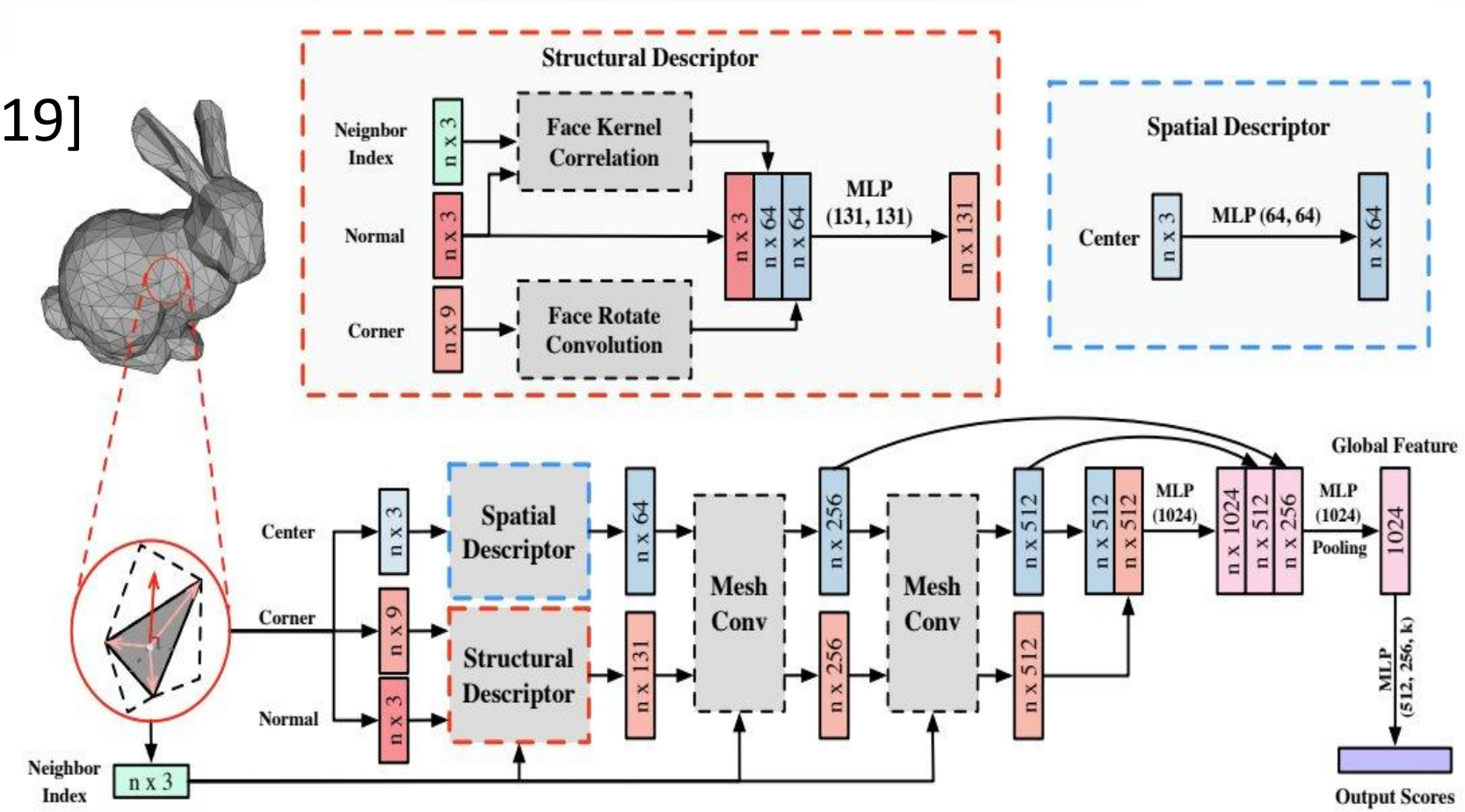
- Retrieval Type
  - Unimodal
  - Cross-modal
  - Multimodal



SOTA Approaches (2) [Gezawa20, Qi21]

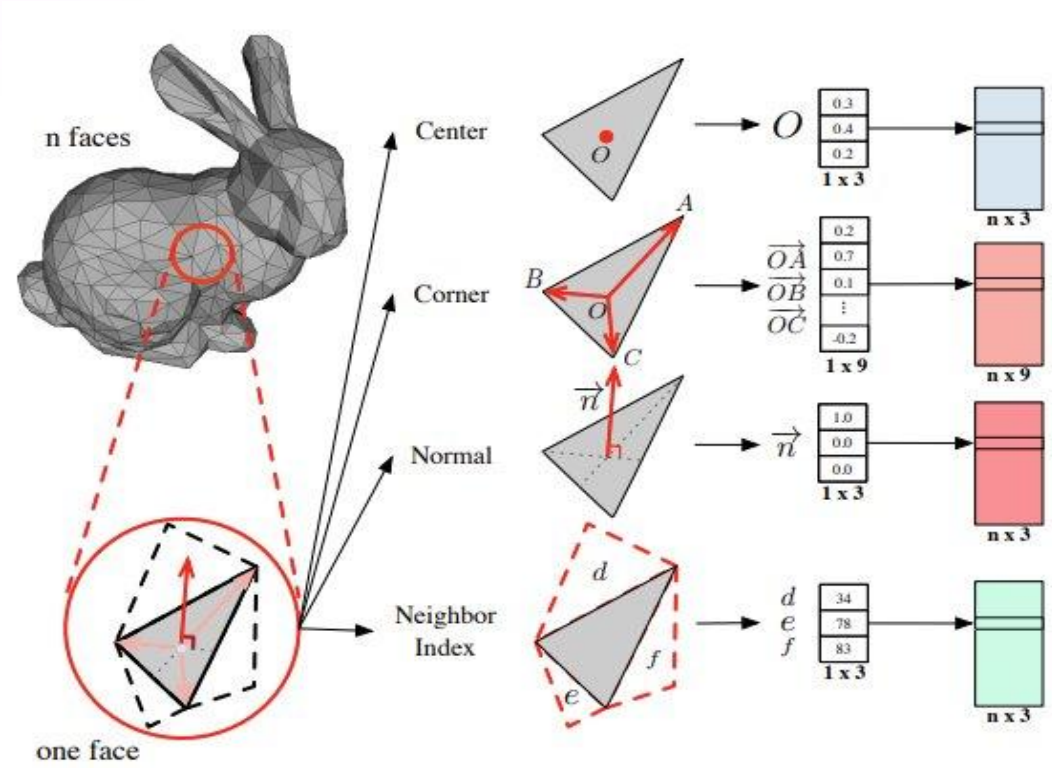
Unimodal Approaches

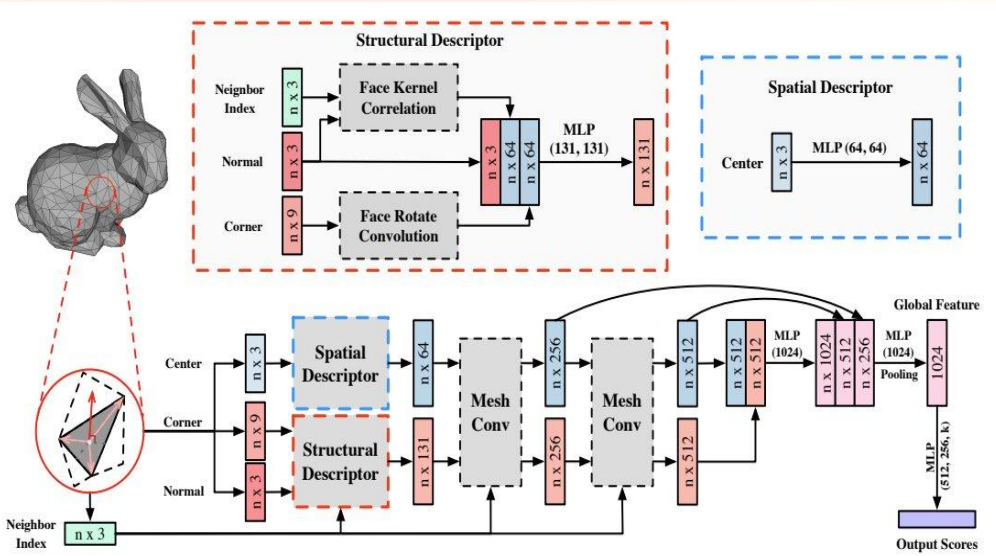
- MeshNet [Feng19]



Unimodal Approaches

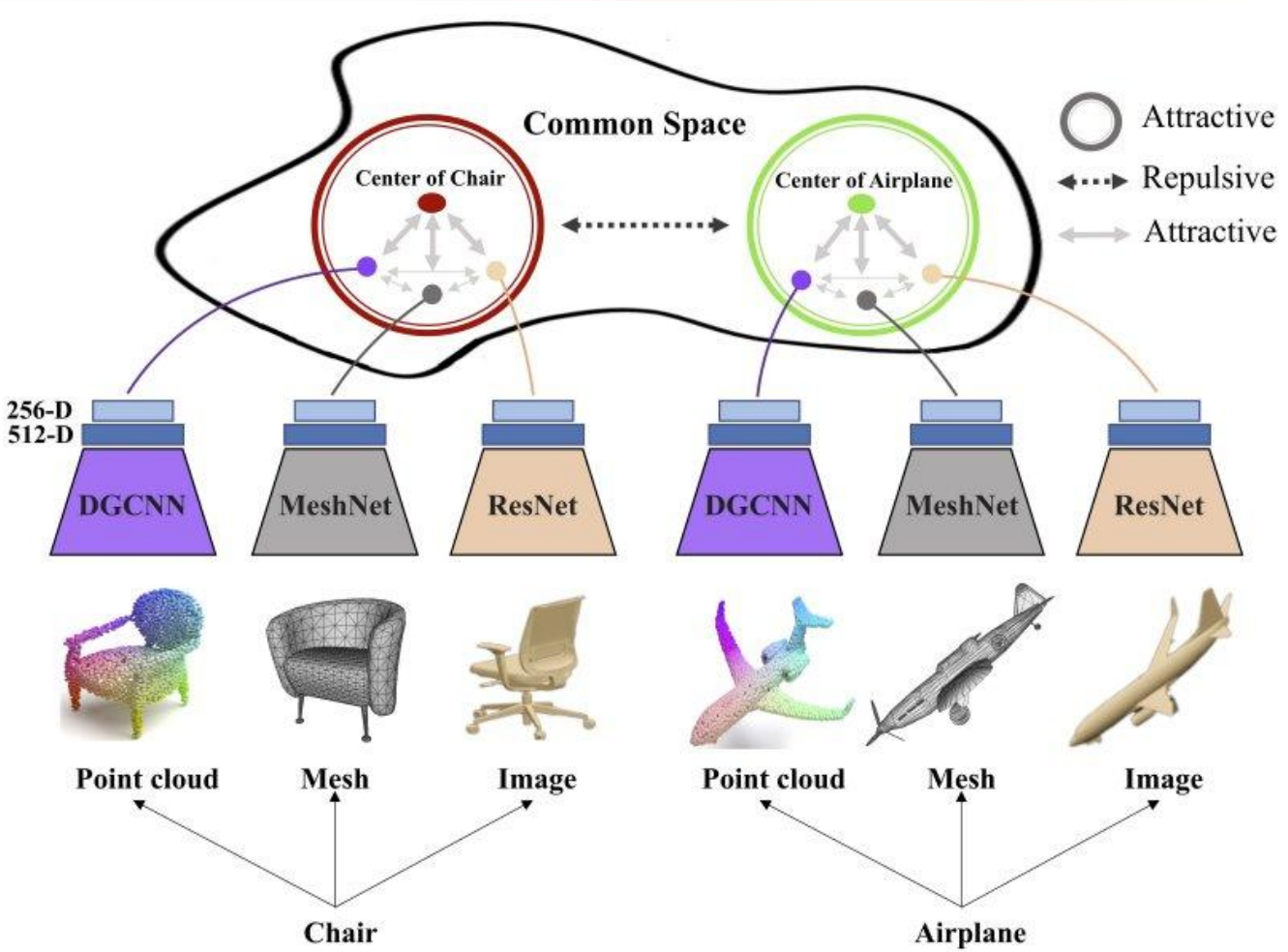
- MeshNet [Feng19]





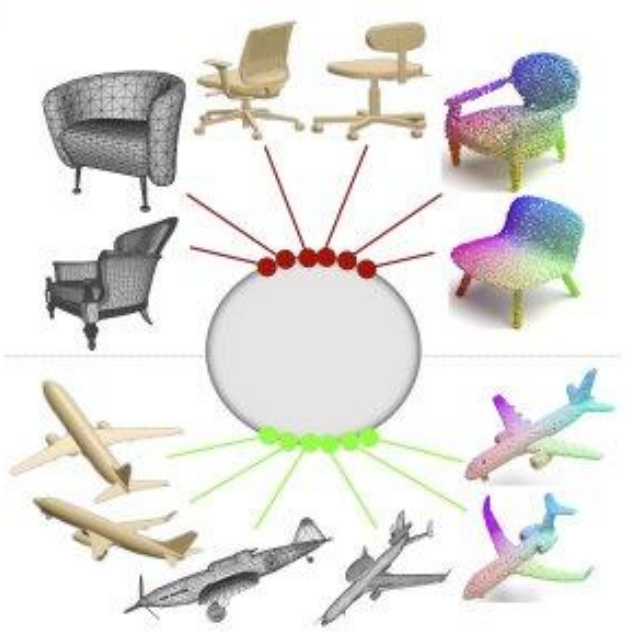
Cross-modal Approaches

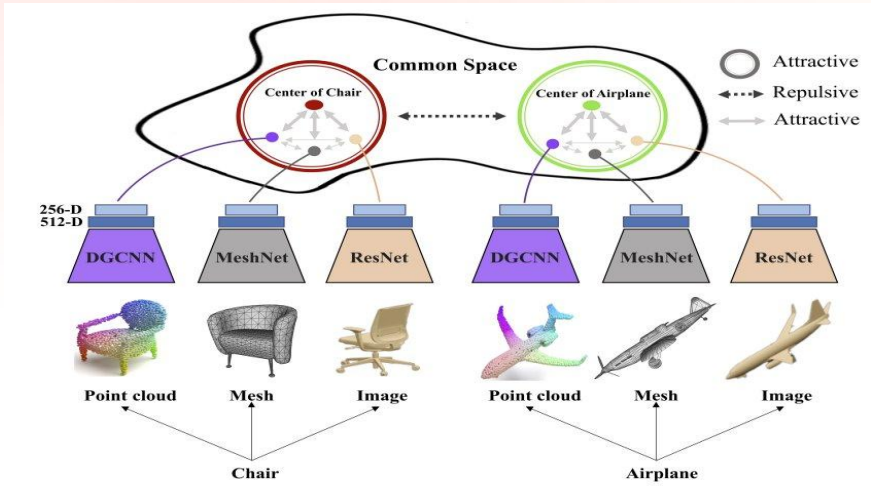
- CMCL [Jing21]

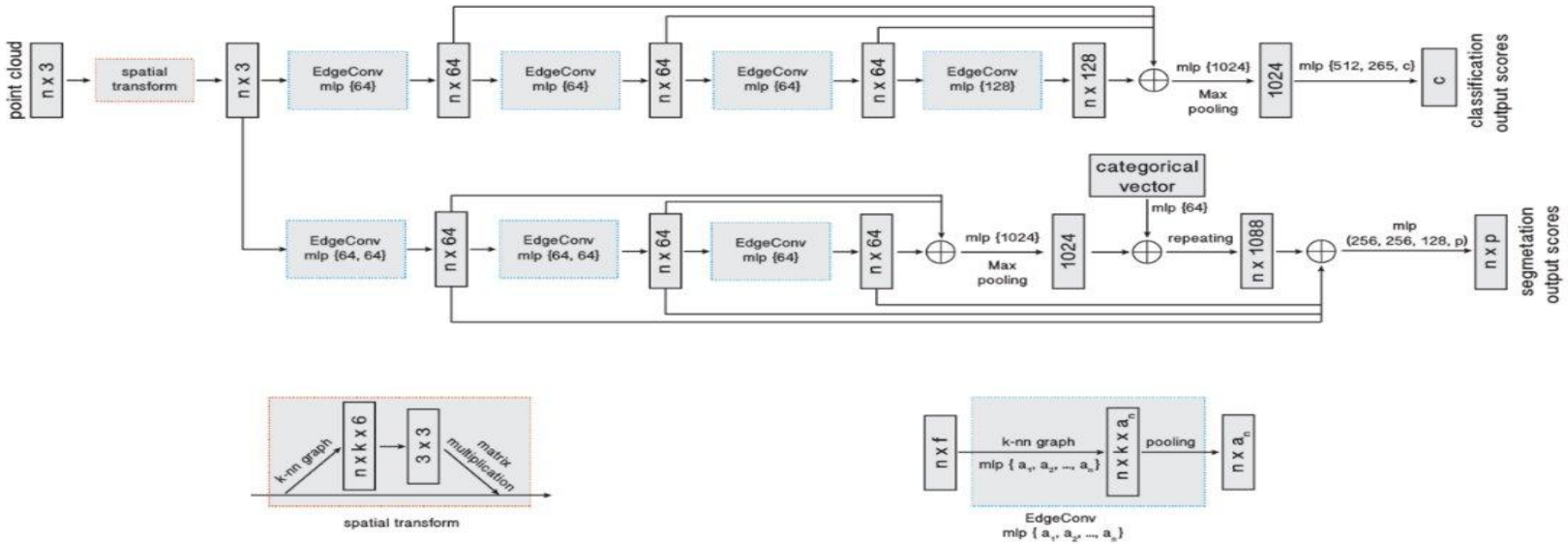


Cross-modal Approaches

- CMCL [Jing21]

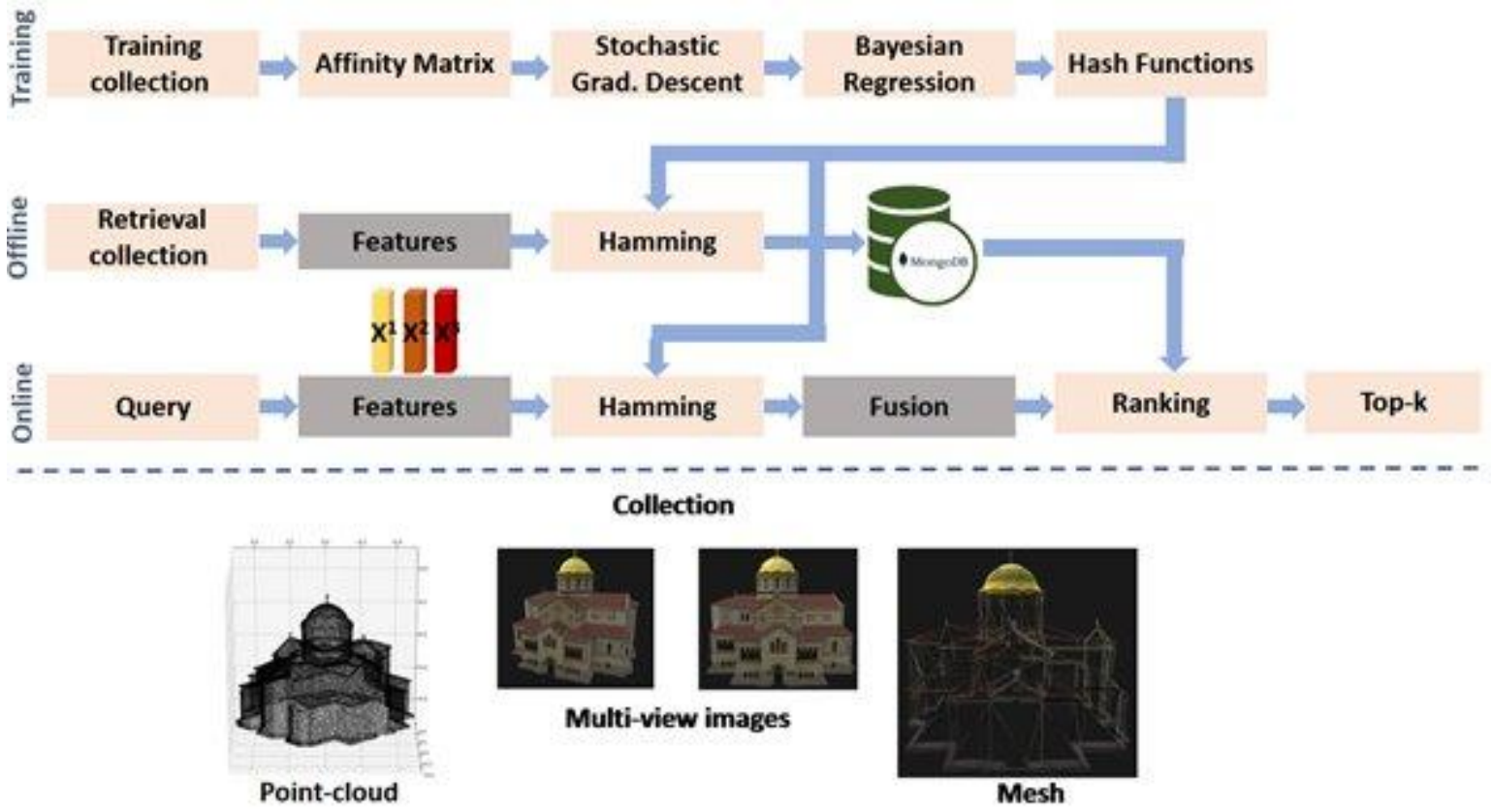






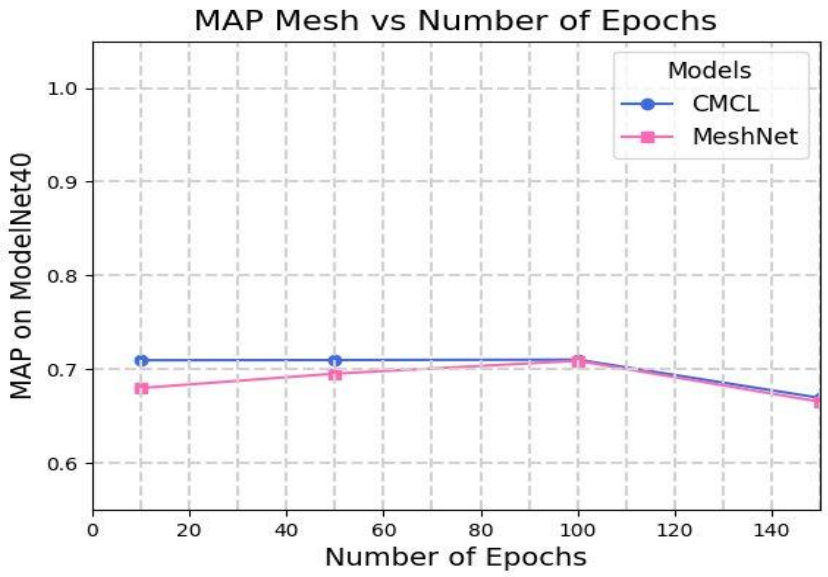
Multimodal Approaches

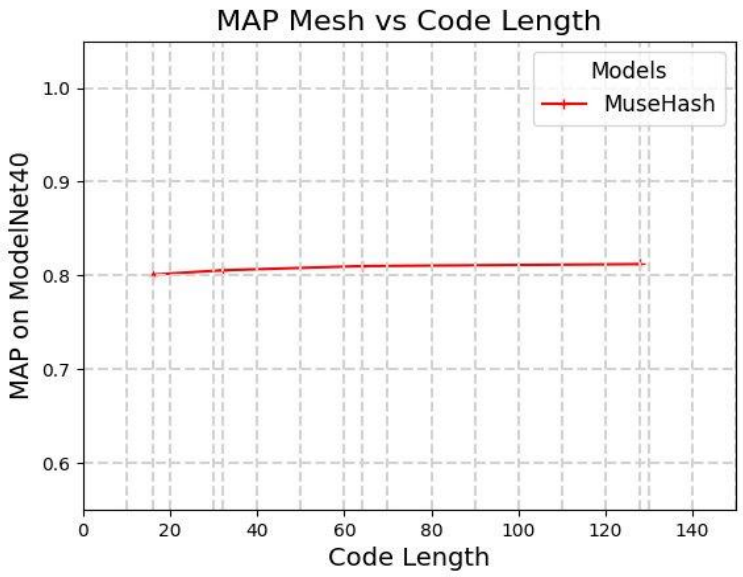
- MuseHash [Pegia24]



Comparison

- MeshNet, CMCL: DNNs for 3D retrieval
- MuseHash: Multimodal hashing for image retrieval
  - Faster queries, Lower needs of memory, Longer training time





SOTA Approaches (3)

- Other interesting works
  - Unimodal: MeshCNN, PointNet, etc.
  - Cross-modal: CMIC
  - Multimodal: LAH, SCA-PVNet, etc.

Future Steps

- Evaluation Benchmark Development
- Generative Models
- Deep Multimodal Fusion Models
- Hybrid Approaches
- Influence of Combining Modalities

3D Retrieval and Query-by-Sketch

- Challenge: create query objects (sketches)
- Idea: “virtual sculpting” in VR [B20]
- Apply Constructive Solid Geometry (CSG)
  - Create shapes by composing simple primitive shapes (Cubes, Spheres, Cylinders, etc.)
  - Two main operations: union and difference

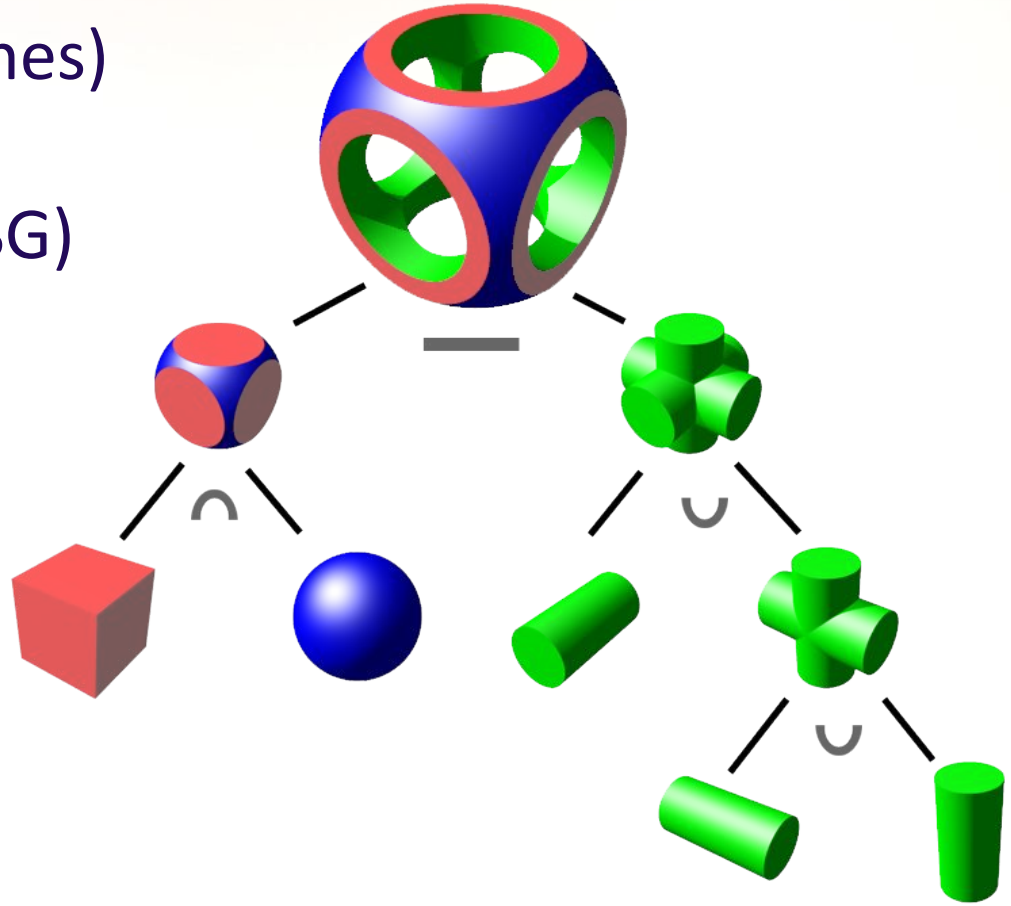
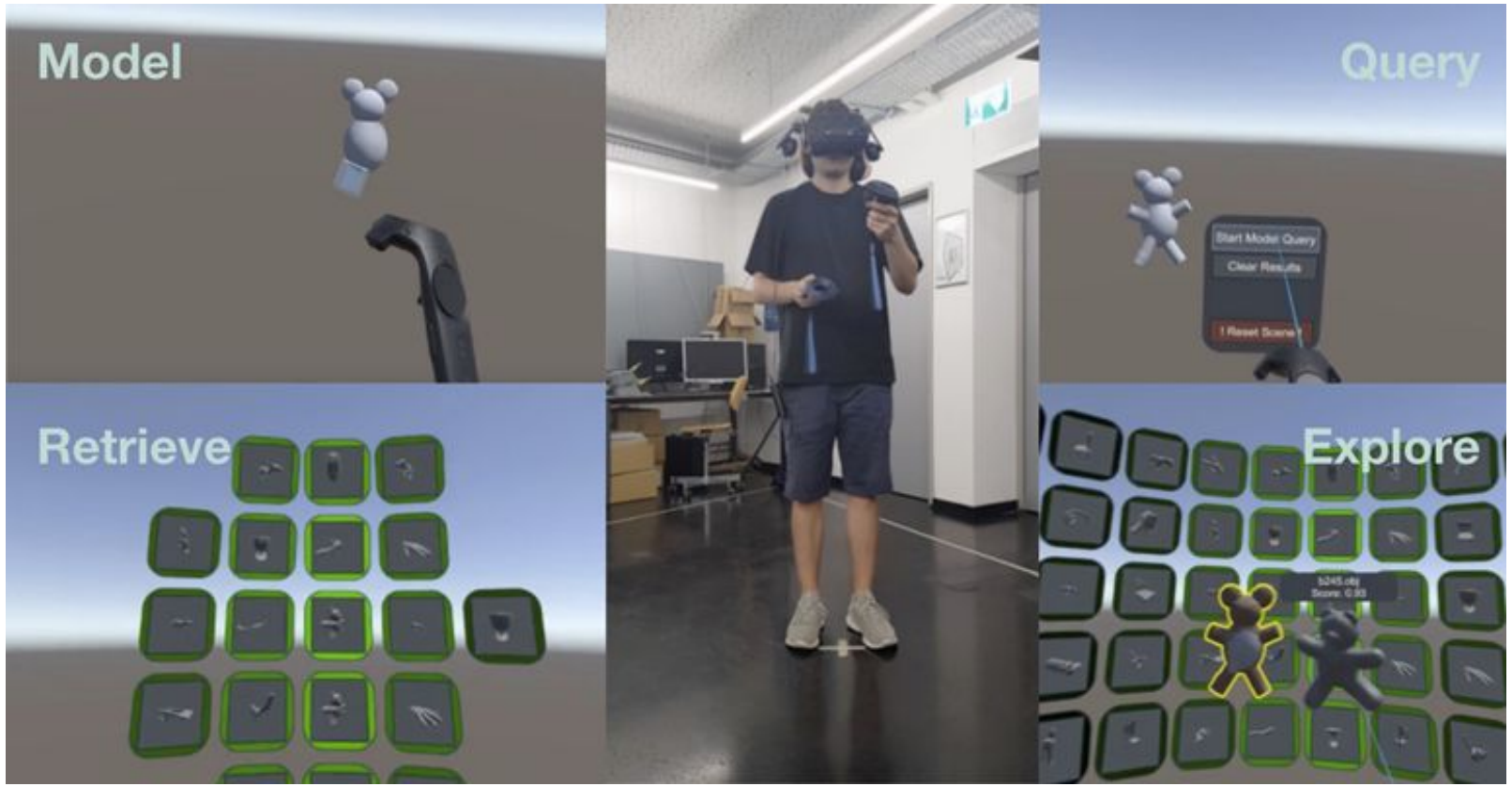


Image source: https://en.wikipedia.org/wiki/Constructive_solid_geometry

CSG for Creating 3D sketches – Demo Video



Evaluating XR Search and Exploration

Image from: https://videobrowsershowdown.org/

Evaluation of Interactive Video Retrieval

- Interfaces are inherently developed for human users
- Every user might be different
  - Different culture, knowledge, preferences, experiences, ...
  - Even the same user at a different time
- Video search interfaces need to be evaluated with real users...
  - No simulations!
  - User studies and campaigns (TRECVID, MediaEval, VBS, LSC)!
  - Find out how well users perform with a specific system
- ...and with real data!
  - Real videos “in the wild” (e.g., IACC.1 and V3C dataset)
  - Actual queries that would make sense in practice
  - Comparable evaluations (same data, same conditions, etc.)

Overview of Evaluation Approaches

- Qualitative user study/survey
  - Self report: ask users about their experience with the tool, thinking aloud tests, etc.
  - Using psychophysiological measurements (e.g., electrodermal activity - EDA)
- Log-file analysis
  - Analyze server and/or client-side interaction patterns
  - Measure time needed for certain actions, etc.
- Question answering
  - Ask questions about content (open, multiple choice) to assess which content users found
- Indirect/task-based evaluation (Cranfield paradigm)
  - Pose certain tasks, measure the effectiveness of solving the task
  - Quantitative user study with many users and trials
  - Open competition, as in VBS, LSC, and TRECVID

Properties of Evaluation Approaches

- Availability and level of detail of ground truth
  - None (e.g., questionnaires, logs)
  - Detailed and complete (e.g., retrieval tasks)
- Effort during experiments
  - Low (automatic check against ground truth)
  - Moderate (answers need to checked by human, e.g. live judges)
  - High (observation of or interview with participants)
- Controlled conditions
  - All users in same room with same setup (typical user-study)
    vs. participants via online survey
- Statistical tests!
  - We can only conclude that one interactive tool is better than the other, if there is statistically significant proof
    - Tests like ANOVA, t-tests, Wilcoxon-signed rank tests, …
    - Consider prerequisites of specific test

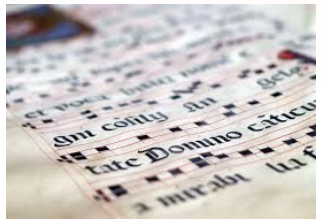
Practicality of Evaluation Approaches

- Task-based approaches have a number of practical advantages
  - ground truth can be reused
  - ground truth can be defined before (in theory, with limits …)
  - automatic assessment (if we have all the ground truth)
  - thus repeatable
  - objective results to chosen level of detail
- Approaches are in principle applicable to XR-based approaches
  - ensure that surveys/question answering do not interfere with immersive experience
  - VR can be considered a user interface like others
  - AR is more challenging, requires controlling the conditions in the real world for comparable results, or adjusting for differences in the real environment

Task Types: Introduction

- Searching for content can be modelled as different types of tasks
- Task serves as a laboratory model of a real-world situation with an information need for multimedia data
  - Isolate: consider one or few steps from a process
  - Standardise: create a framework of controlled conditions (applicability of evaluation metrics, repeatability)
  - Simplify: reduce complexity of the setting: to limit the number of variables
  - Task design choices impact dataset preparation, annotations (effort!), evaluation methods and the way to run the experiments

Task Types: Overview



Example

Visual cue
Aural cue

Music scores

Textual

None

Query

Specific item

Result set

Class

Target

KIS by example

KIS

QbE

Retrieval

Exploration

Task Types: Overview

A_{CI}: Number of correct items satisfying a search need in a dataset

A_{SI}: Requested number of submitted correct items

A_{PM}: Search need presentation modality

A_{PT}: Search need presentation timing

A_{PQ}: Search need presentation quality

A_{DC}: Data collection

A_{TL}: Time limit

A_{US}: User skills

A_{NU}: Number of operating users

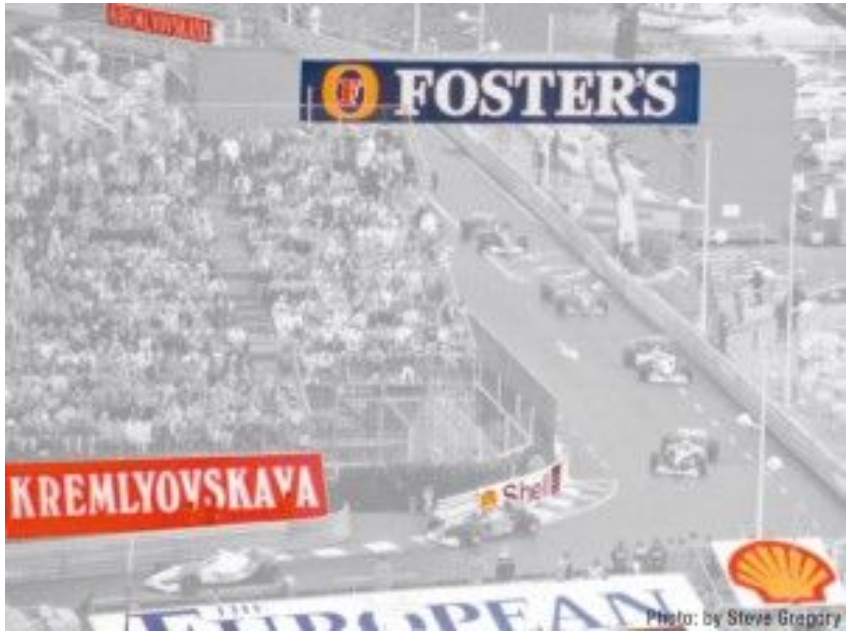
A_{QM}: Quality measure

Table 2. VBS'21 task categories represented as vectors of the task space. For each category, the value in an axis column presents the currently used axis option (specified in Section 3). Due to the virtual conference setting, only expert users were participating.

Task Name	A_{CI}	A_{SI}	A_{PM}	A_{PT}	A_{PQ}	A_{DC}	A_{TL}	A_{US}	A_{NU}	A_{QM}
Visual KIS	1	1	1	2	1	1	1	1	2	1, 3
Textual KIS	1	1	2	3	2	1	1	1	2	1, 3
Ad-hoc search	3	3	2	2	2	1	1	1	2	2, 3

Task Types: Query by Example

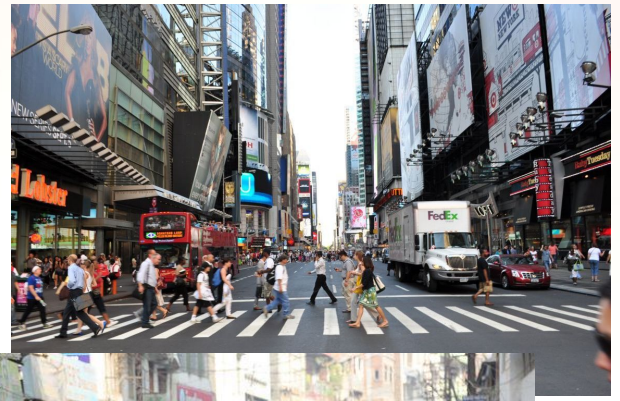
- User holds a digital representation of a relevant example of the needed information
- Example or its features can be sent to system
- User does not need to translate example into query representation
- e.g., trademark/logo detection



Task Types: Known Item Search (KIS)

- User sees/hears/reads a representation of the needed information
  - Used in VBS & LSC
- Representation of exactly **one** relevant item/segment in content set
- Models cases where the user has a memory of content to be found
- User must translate the representation to query methods supported by the system
  - The complexity of this translation depends significantly on the modality
  - e.g., visual is usually easier than textual, which leaves more room for interpretation
  - Relation of/to content is important
  - e.g. searching in own life log media vs. searching in media collection on the web

“on a busy street”

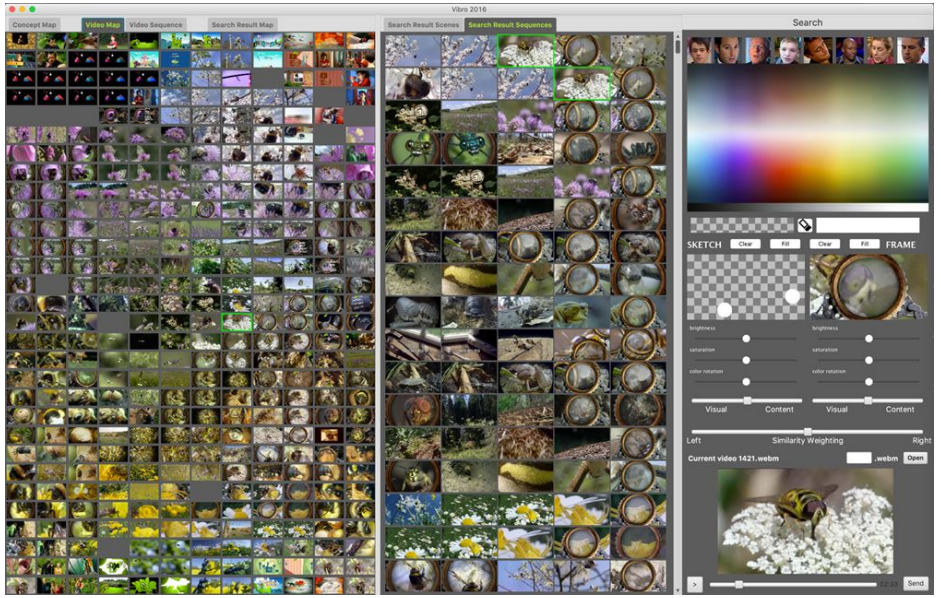


Task Types: Retrieval

- User sees/hears/reads a representation of the needed information
- Representation of a broader set/class of relevant items/segments
  - cf. TRECVID AVS task
- Models cases where the user has a memory of the type of relevant content
- Similar issues of translating the representation like for KIS, but due to broader set of relevant items the correct interpretation of textual information is a less critical issue
- Raises issues of what is considered within/without scope of a result set
  - e.g., partly visible, visible on a screen in the content, cartoon/drawing versions, …
  - TRECVID has developed guidelines for annotation of ground truth

Task Types: Exploration

- User does not start from a clear idea of the information need
- Browsing and exploring may lead to identifying useful content
- Reflects a number of practical situations, but very hard to evaluate
- No known examples of such tasks in benchmarking campaigns due to the difficulties with evaluation



Demo:
https://www.picsbuffet.com/

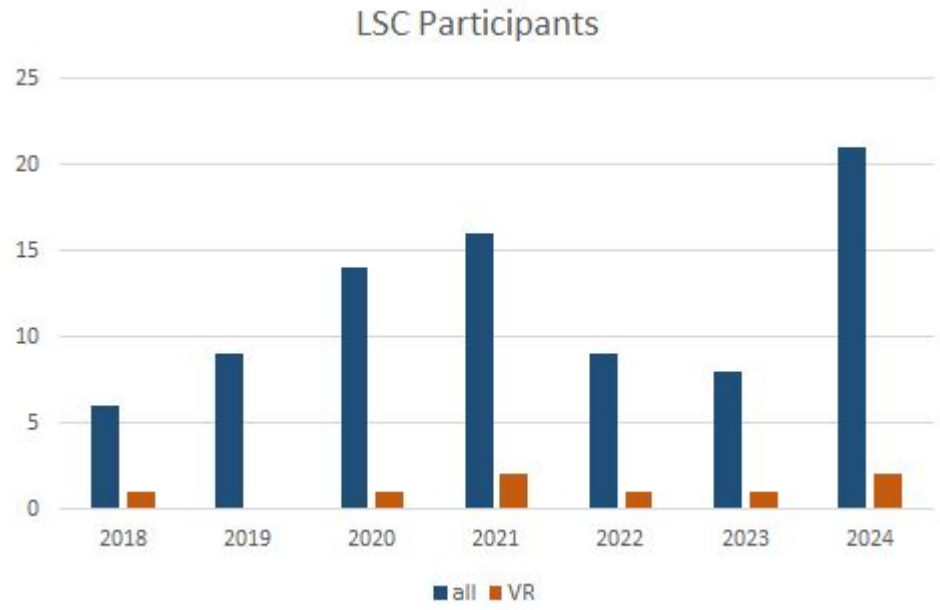
Task design is about trade-offs: Aspects to consider

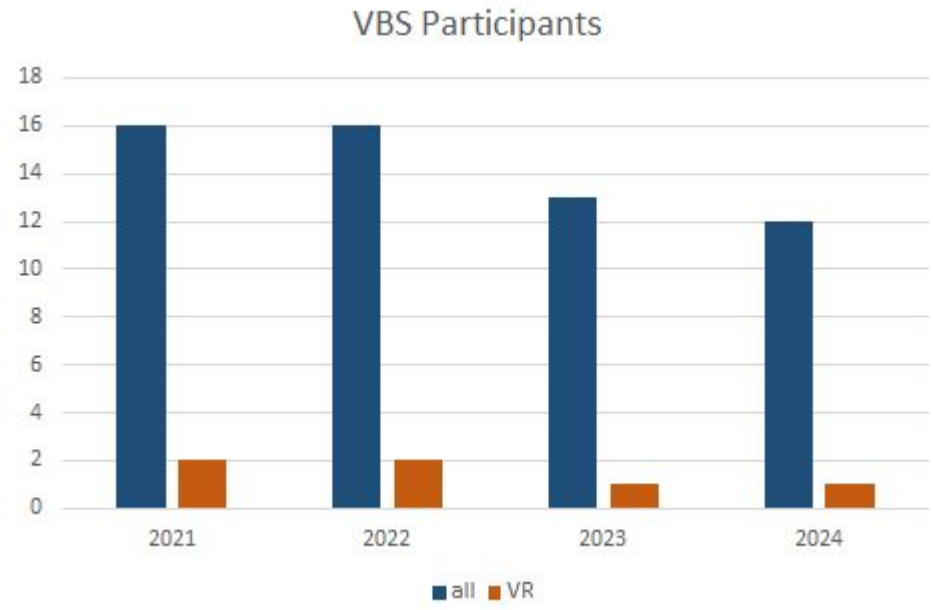
- Tasks shall
  - model real-world content search problems, in order to assess whether tools are usable for these problems
  - set controlled conditions, to enable reliable assessment
  - be repeatable, to compare results from different evaluation sessions
  - avoid bias towards certain features or query methods

many real world problems involve very fuzzy information needs	well defined queries are best suited for evaluation
users remember more about the scene when they start looking through examples	information in the task should be provided at defined points in time
during evaluation sessions, relevant shots may be discovered, and the ground truth updated	for repeatable evaluation, a fixed ground truth set is desirable
although real world tasks may involve time pressure, it would be best to measure the time until the task is solved	time limits are needed in evaluation sessions for practical reasons

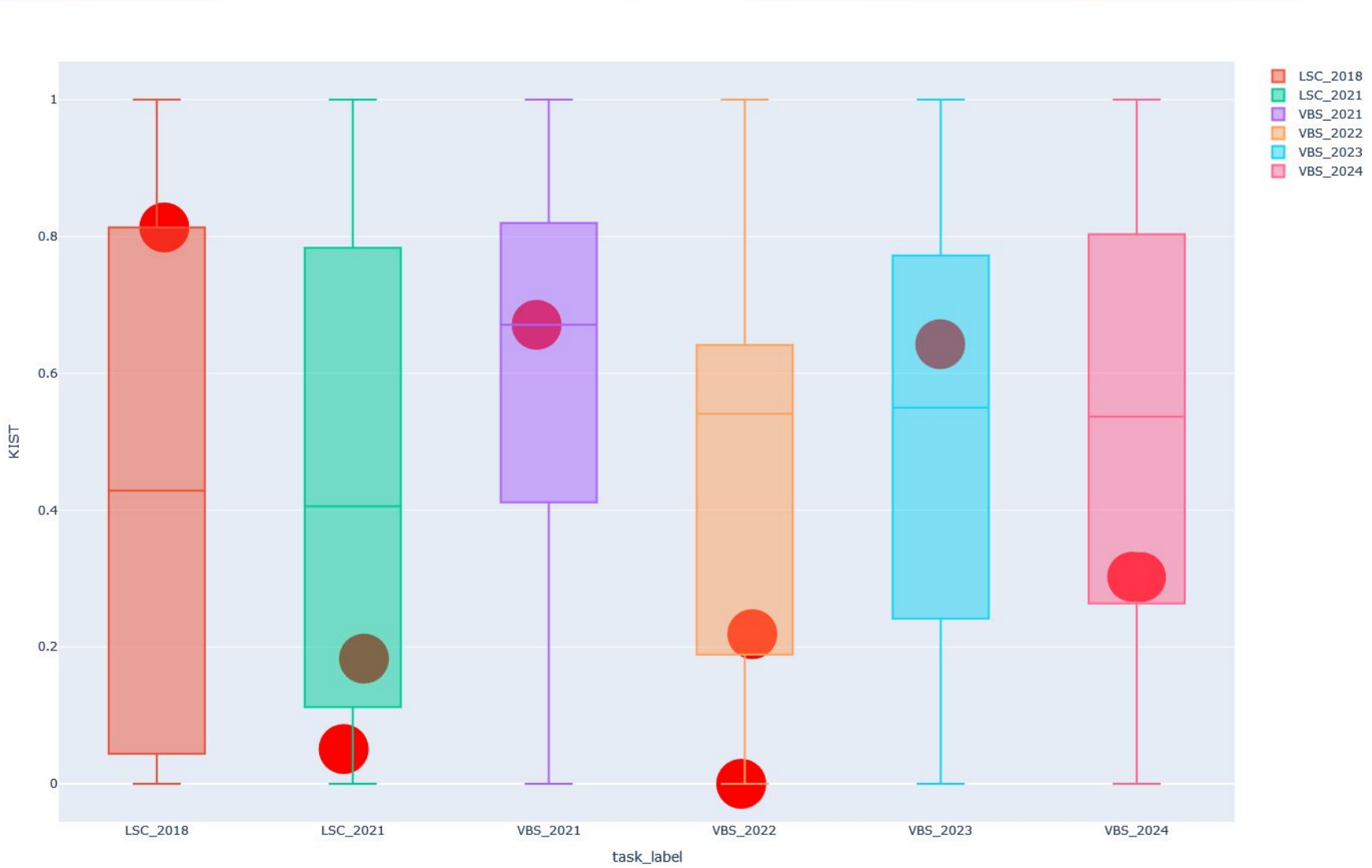
XR systems in evaluation campaigns

- XR always means VR in this case
- we looked into VBS, LSC, NTCIR and TRECVID
- NTCIR: one VR system for exploratory task in 2017
- TRECVID: one VR system for Instance Search in 2017

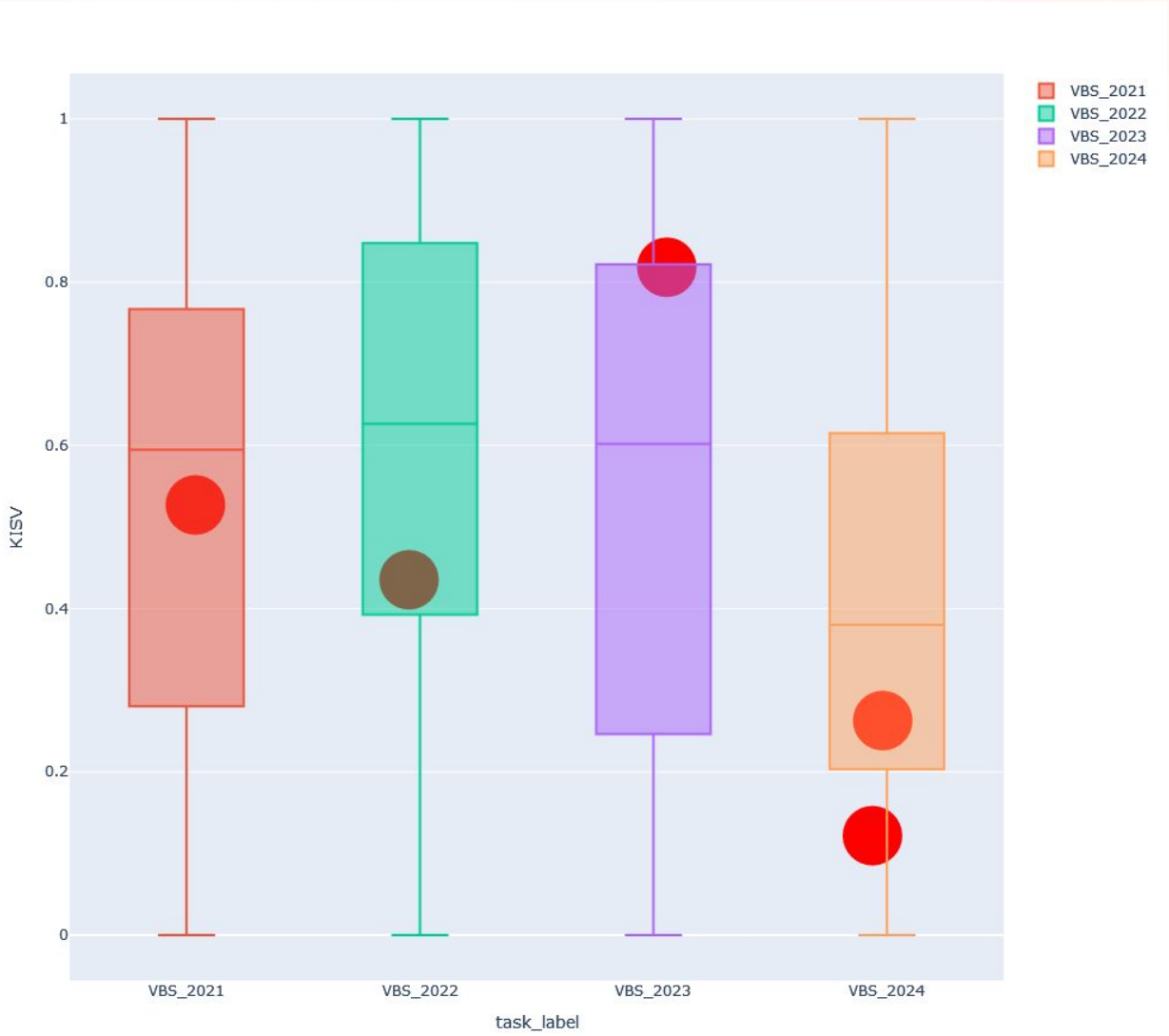




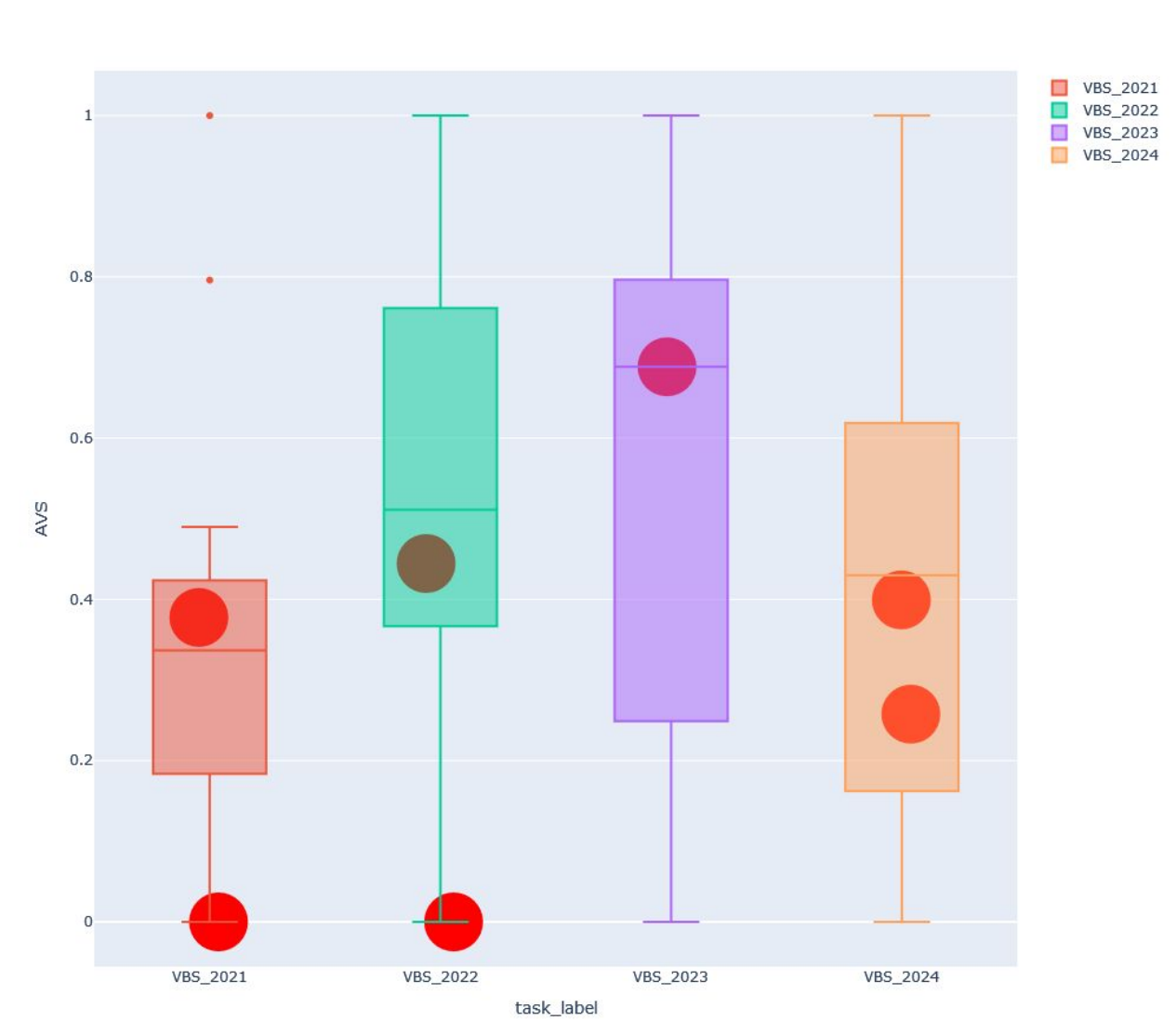
Relative scores of VR systems: Textual KIS



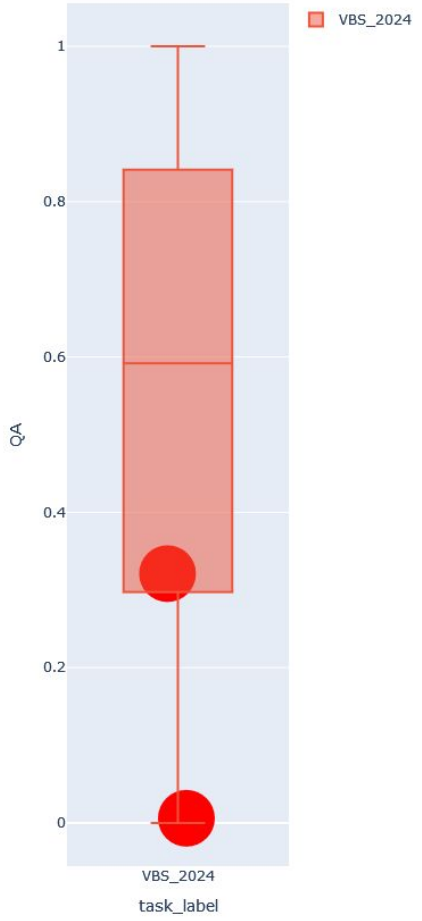
Relative scores of VR systems: Visual KIS

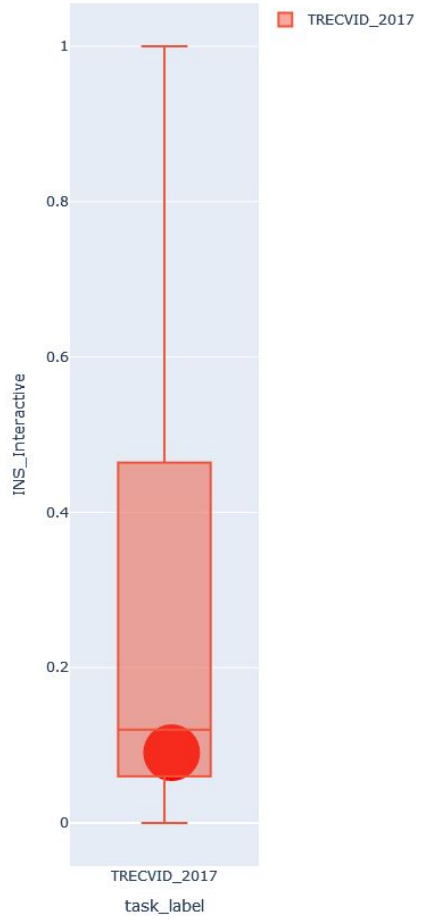


Relative scores of VR systems: AVS



Relative scores of VR systems: QA and INS

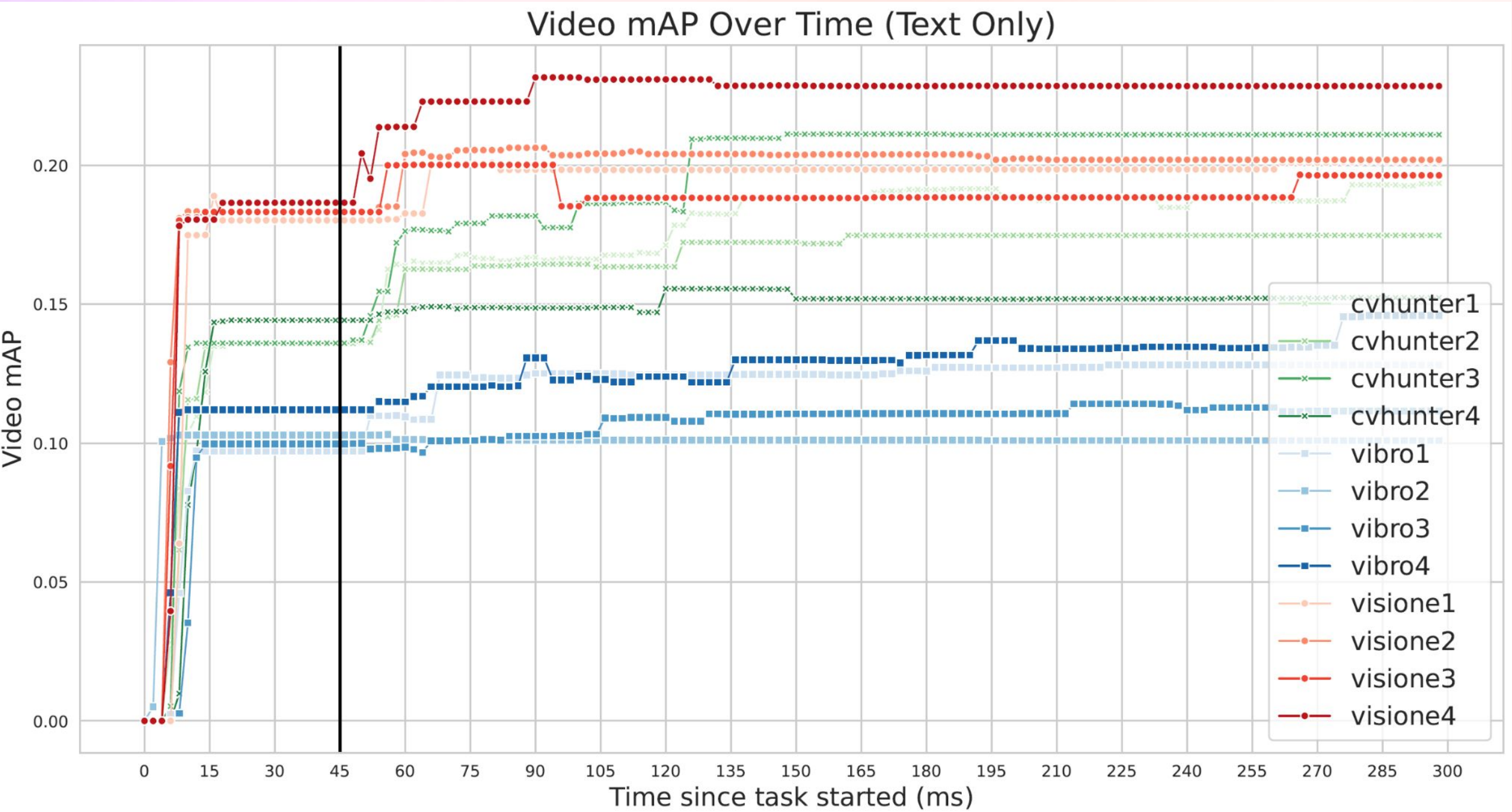


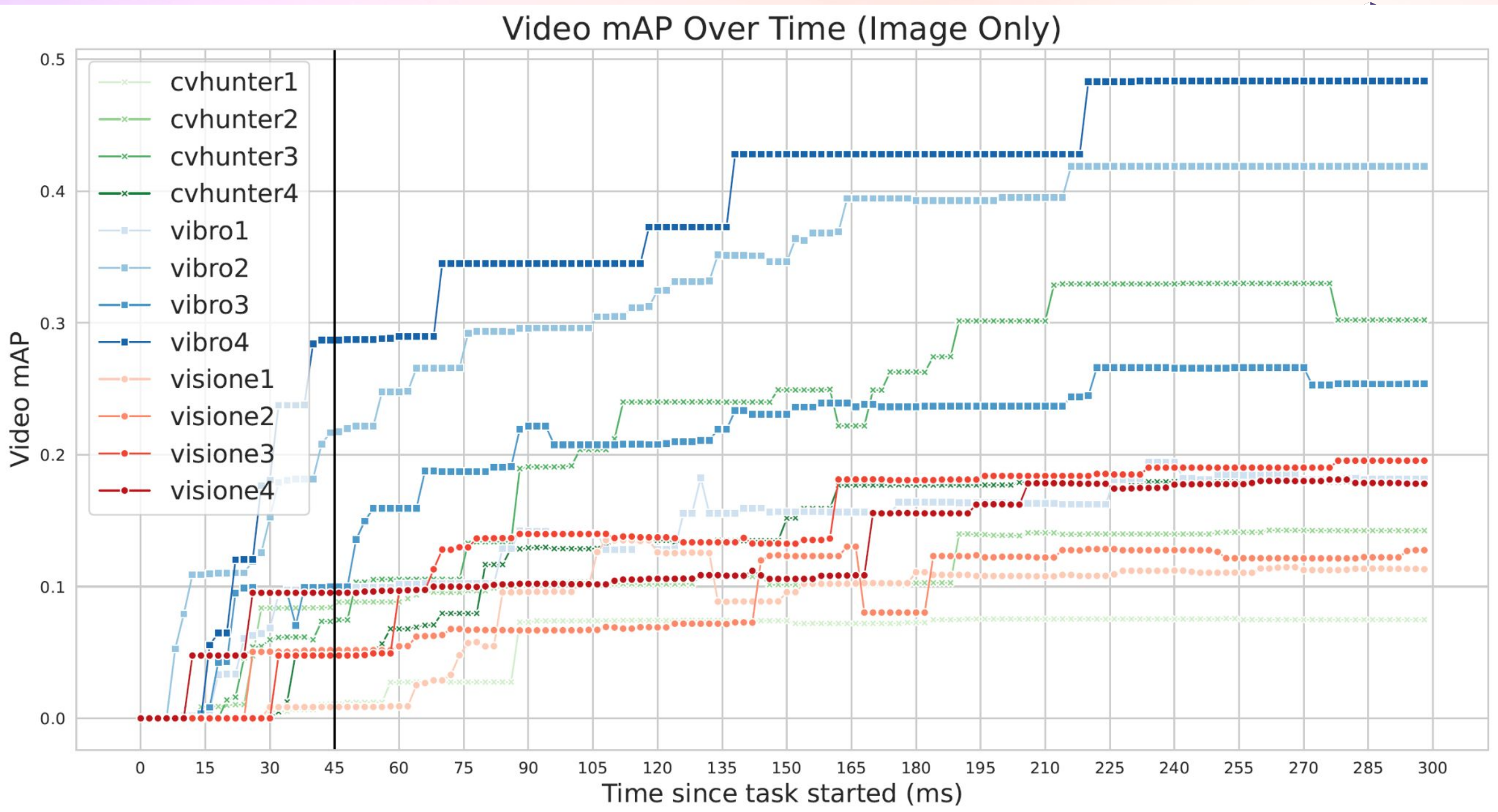


The impact of vision-language embeddings

- Successful interactive video retrieval systems rely strongly on vision-language embeddings (e.g. CLIP) [S24]
- These models enable fast text search
- For queries where the embedding works well, and the number of items is not too large, the initial text query might solve the task
- For harder tasks, other query modalities and exploration have the potential to significantly enhance results

Times of query changes and submission per task and team


Video mAP Over Time (Text Only)
Video mAP
0.00
0.05
0.10
0.15
0.20
0
15
30
45
60
75
90
105
120
135
150
165
180
195
210
225
240
255
270
285
300
Time since task started (ms)
cvhunter1
cvhunter2
cvhunter3
cvhunter4
vibro1
vibro2
vibro3
vibro4
visione1
visione2
visione3
visione4


Video mAP Over Time (Image Only)
cvhunter1
cvhunter2
cvhunter3
cvhunter4
vibro1
vibro2
vibro3
vibro4
visione1
visione2
visione3
visione4
Video mAP
0.0
0.1
0.2
0.3
0.4
0.5
0
15
30
45
60
75
90
105
120
135
150
165
180
195
210
225
240
255
270
285
300
Time since task started (ms)

Desktop Search vs. Search in VR …

Comparison of user interaction in multimedia retrieval at VBS 2023 [S23b]

- Participation of vitrivr and vitrivr-VR …
- … using the same dataset (VSC1 and V3C2) [R18]
- … using the same database (Cottontail) [G20]
- … using the same retrieval engine (Cineast) [R16]

- Only difference was the user interaction (desktop vs. VR)

… Desktop Search vs. Search in VR …

Measures

- **Complexity of query formulation**: Time to first results
- **Lowest rank of a target item**: Result quality
- **Browsing performance**: Browsing miss@k
- **Find object in result set**: Relation between best rank & browsing time

… Desktop Search vs. Search in VR …

Overall Statistics

Task Type	Points	Tasks solved	Correct Submissions	Incorrect Submissions
V-KIS	738	4 / 6	4	0
	877	5 / 6	5	3
V-KIS M	618	4 / 6	4	1
	881	6 / 6	6	0
T-KIS	**928**	6 / 7	7[1]	1
	738	5 / 7	5	2
AVS	**702**	7 / 7	231	82
	699	7 / 7	236	109
Total	2986	21 / 26	246	84
	3195	**23** / 26	252	114

Source: [S23b]

Median time from task start to first explorable results

Task Type	Modality	Median Time to First Result in Seconds	Difference
V-KIS	Desktop	**30.89**	+26.9%
	VR	39.20	
V-KIS M	Desktop	**27.43**	+17.2%
	VR	32.15	
T-KIS	Desktop	**21.29**	+24.7%
	VR	26.55	
AVS	Desktop	**25.41**	+11.98%
	VR	28.45	
Any	Desktop	**25.81**	+24.18%
	VR	32.05	

Upper line: vitrivr / lower line: vitrivr-VR

… Desktop Search vs. Search in VR …

Median rank of the first correct item in the result set

Task Type	Modality	Median Best Video Rank	Median Best Segment Rank
V-KIS	Desktop	**30.5**	n/a
	VR	34.5	n/a
V-KIS M	Desktop	**20**	20
	VR	71.5	74.5
T-KIS	Desktop	**2**	4
	VR	3	n/a
Any	Desktop	**7**	65
	VR	24	1201

Source: [S23b]

Best ranks of target video for all tasks w/o correct submisison

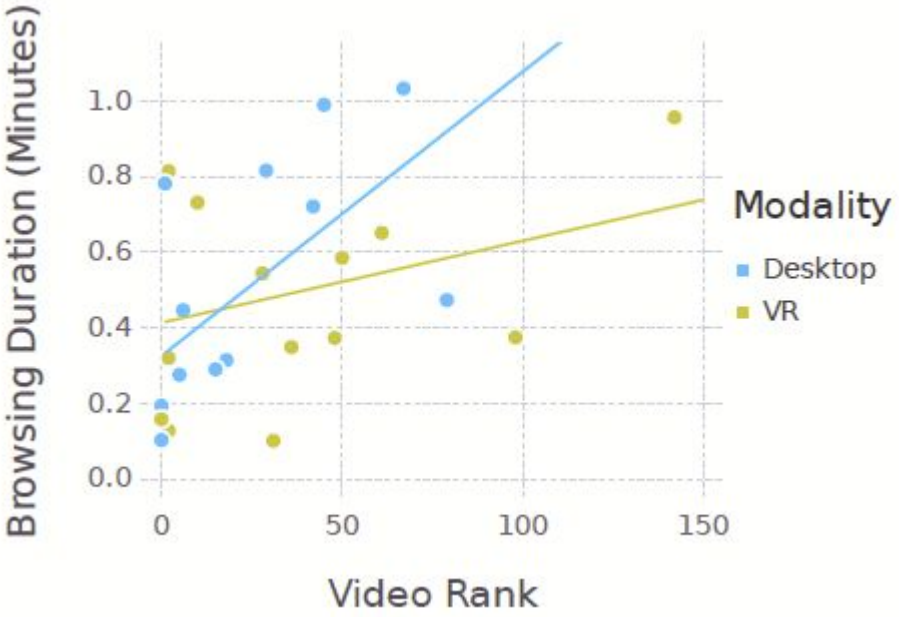
Task Type	Modality	Misses	Min Video Ranks	Min Segment Ranks
V-KIS	Desktop	2	12, 202	n/a, n/a
	VR	**1**	24	n/a
V-KIS M	Desktop	2	24, 641	24, 1613
	VR	**0**	–	-
T-KIS	Desktop	**1**	2	n/a
	VR	2	14, 64	14, n/a

Upper line: vitrivr / lower line: vitrivr-VR

… Desktop Search vs. Search in VR

Results

- VR competition results better than desktop results, even though query formulation took longer
- VR results competition results better than desktop results, even though correct results are ranked lower
- Less browsing time and fewer misses during browsing in VR



- But: more thorough comparison needed

Source: [S23b]

Insights and open questions

- **Caution: small sample of XR systems in benchmarks**
- Quick text entry is useful for some queries
  ⟶ hard in XR
- advances in embeddings brought improvements across systems
  ⟶ XR systems seem to have improved less
- Cranfield paradigm is dominant in benchmarks, open-ended exploration is not well represented due to complexity of evaluation/comparability
  ⟶ task setup is not advantageous for XR systems

Wrap-up

Image from: https://commons.wikimedia.org/wiki/File:Unboxing_Oculus_Go_64GB.jpg

Main take-aways

- XR applications need 3D content
  - can be obtained either by multi-cross modal retrieval for 3D assets (in various representations)
  - can be obtained by search for 2D assets suitable for reconstruction
- XR based search has been around for some time
  - performance in benchmarks still does not surpass that of desktop based systems
  - benchmarks seem to favor systems that are fast with text queries
- Recent advances in XR technologies and ML make it worth to revisit the multimedia search in XR

Research directions

- **Collaborative retrieval**: most existing approaches are tailored to single-user interfaces and do not support collaboration (in query formulation, result analysis, relevance feedback, etc.)
- **Immersion**: address further types of user interactions
- **Evaluation**:
  - Involve more XR systems in benchmarks to better understand their strengths and weaknesses on specific task types, task hardness, …
  - Design tasks that foster exploration while still being comparable and can be evaluated with reasonable effort
  - Design suitable benchmarks for AR-based search systems that leverage the strengths of immersive user interaction


Atos


CERTH
CENTRE FOR
RESEARCH & TECHNOLOGY
HELLAS


Continental


DW
Made for minds.


FFP PRODUCTIONS


FINCONS
FINCONS GROUP


i2cat
THE INTERNET
RESEARCH CENTER


ip rights.gr
Intellectual Property Law & Practice in Greece
Copyright | Designs | Patents | Trademarks


JOANNEUM
RESEARCH
DIGITAL


KU LEUVEN


rrog
DIGITAL MEDIA


NVIDIA
NETWORKING


Rai


Sound


POLITÉCNICA


Unity


University
of Basel


VISYON

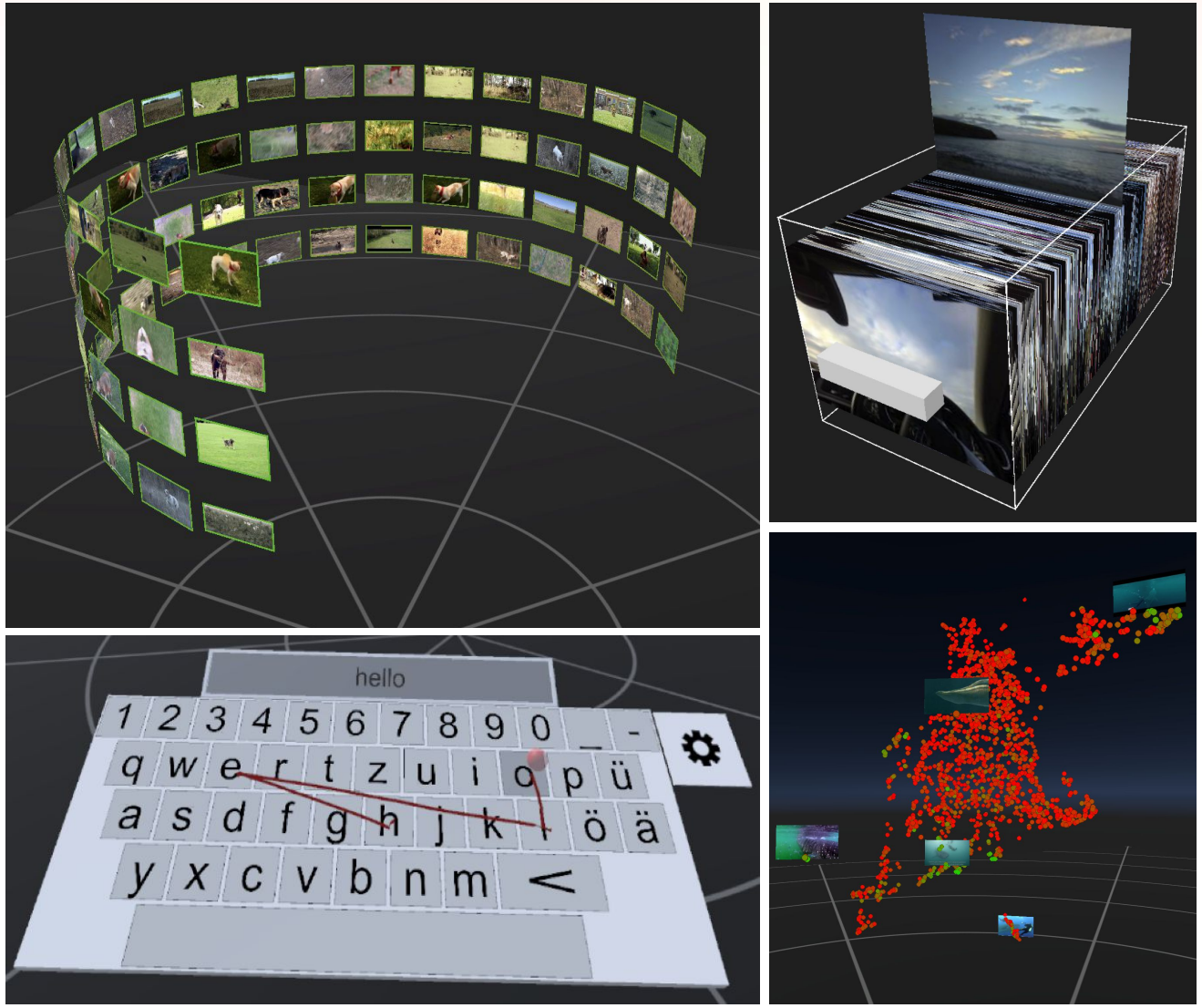

XRB


ZAUBAR

Hands-on: virtrivr-VR

vitrivr-VR

- VR multimedia analytics system
  - Focus on retrieval
- Exploring:
  - Query formulation
  - Interactive visualization
  - Results browsing



User Interaction

1. Grip button to grab and move
2. Trigger button to interact
3. Right controller A button to start speech-to-text (press and hold)
4. Left controller stick rotate cylinder display
5. Left controller X button to open the options menu

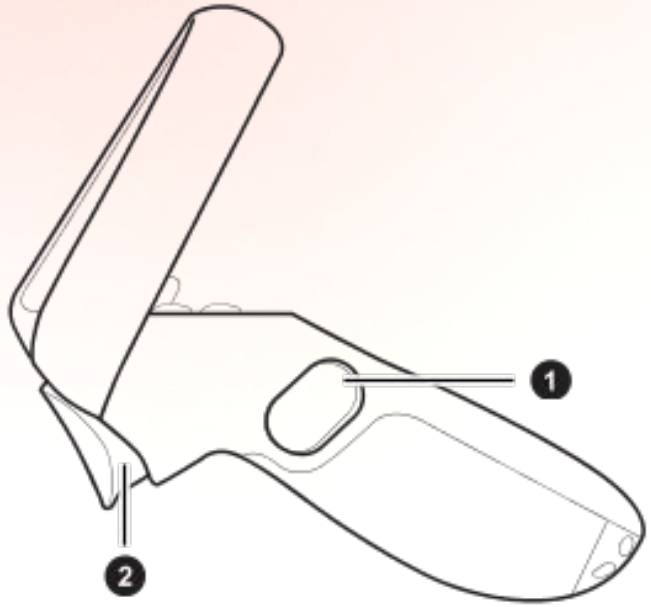
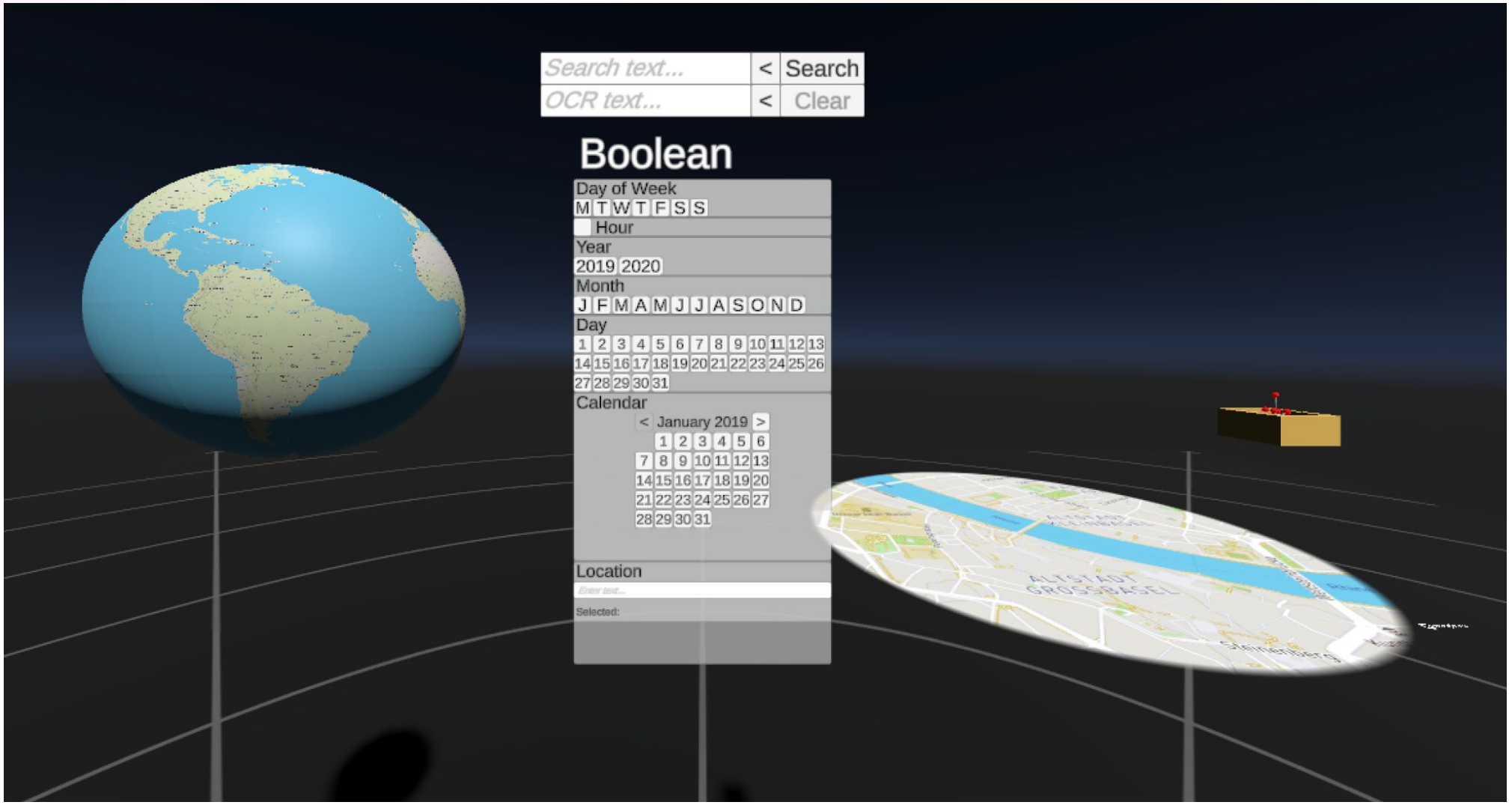
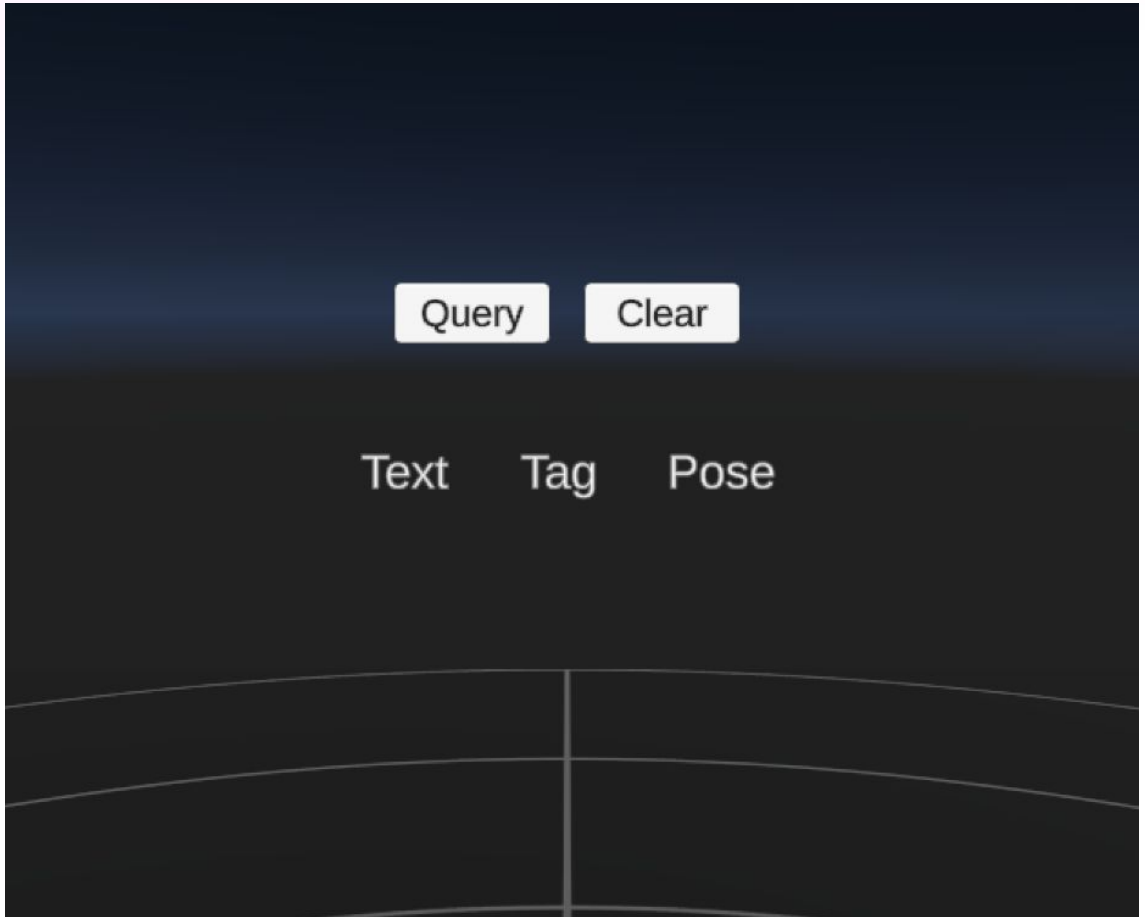


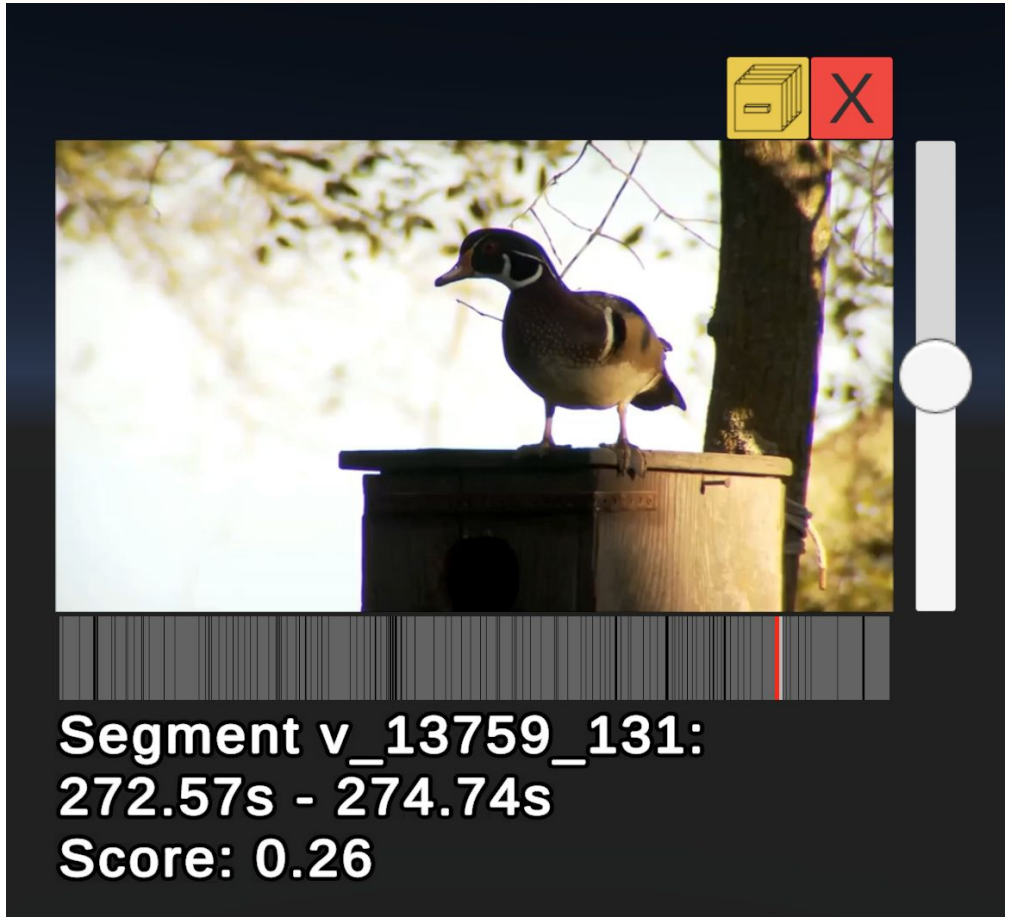
Image source: vive.com

Lifelog Search Interface



Video Search Interface

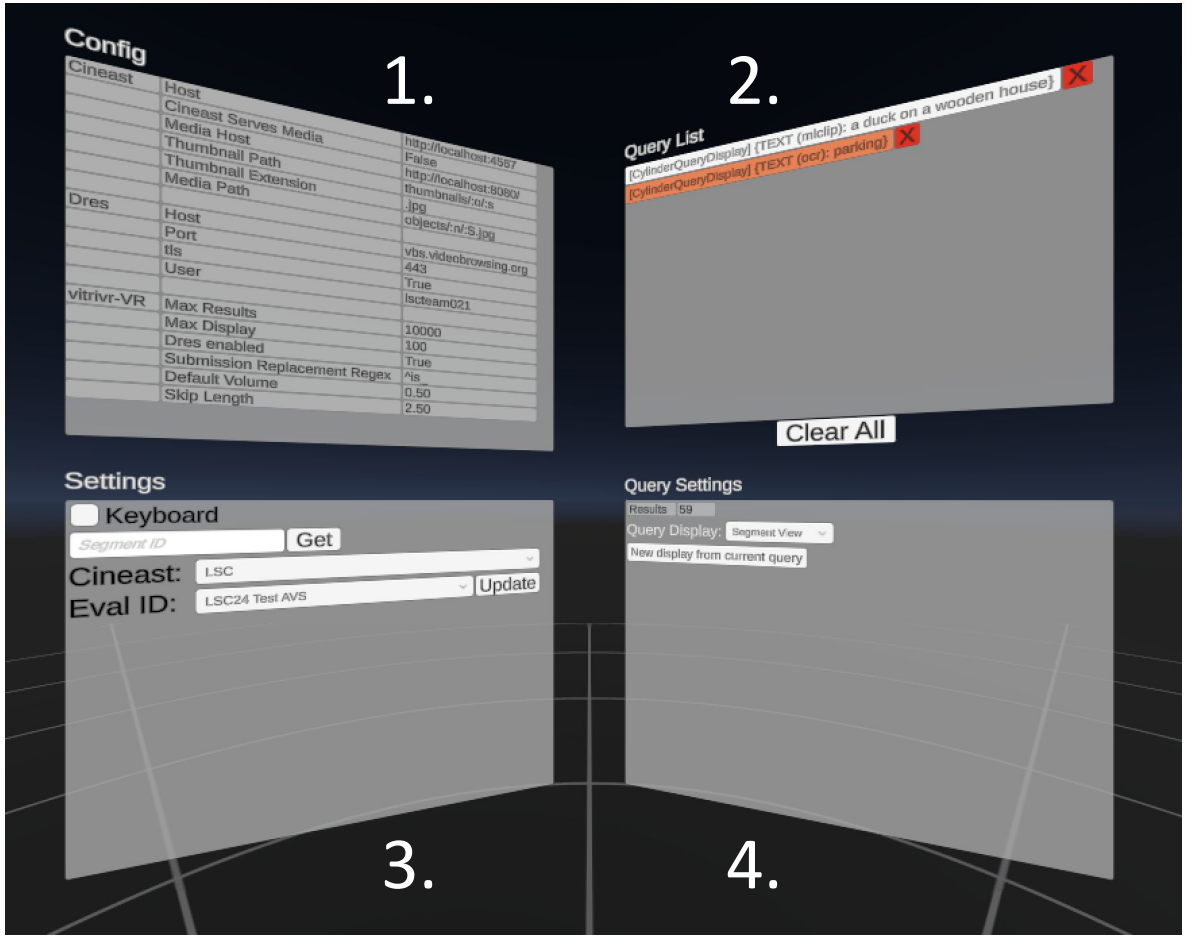




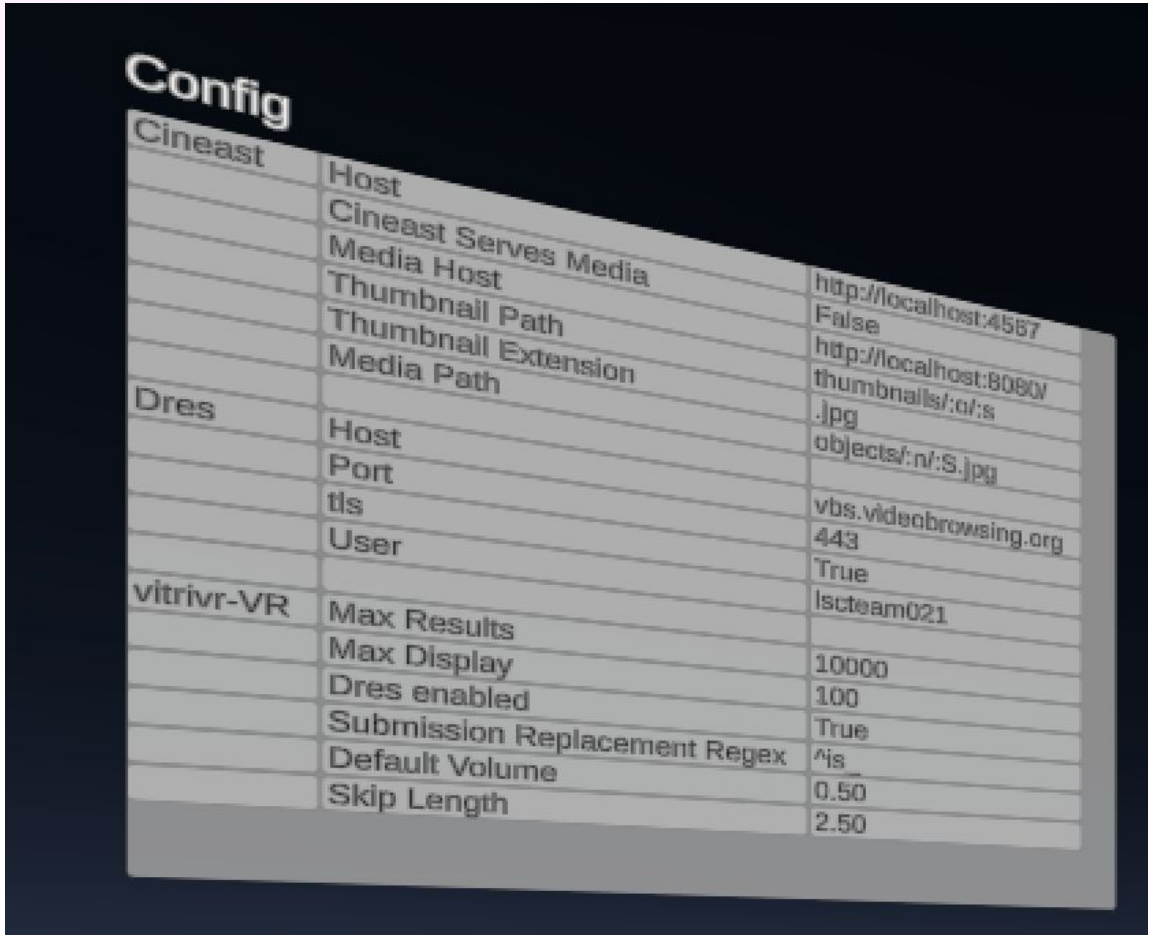
Options Menu

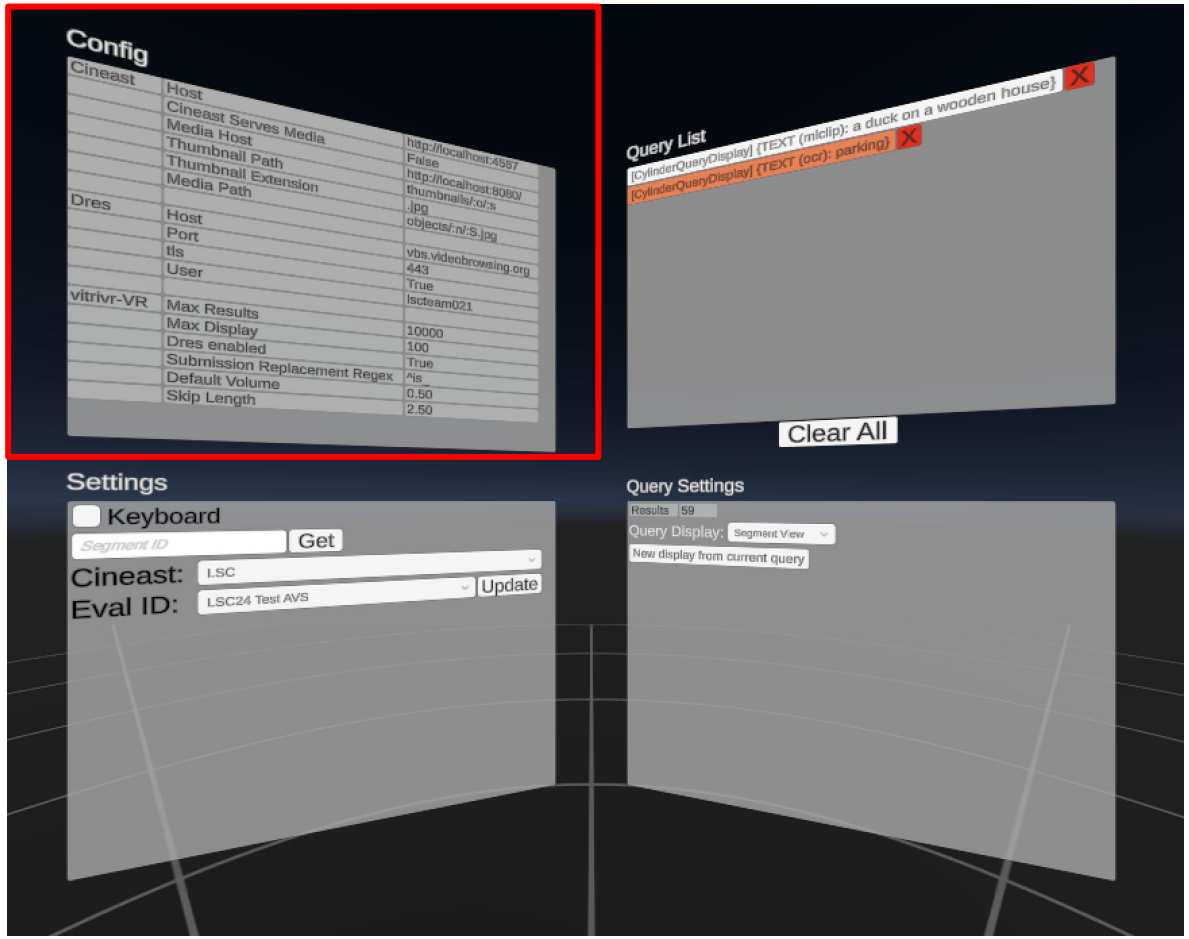
Contains:

1. Configuration information
2. A list of previous queries
3. General settings
4. Query specific settings

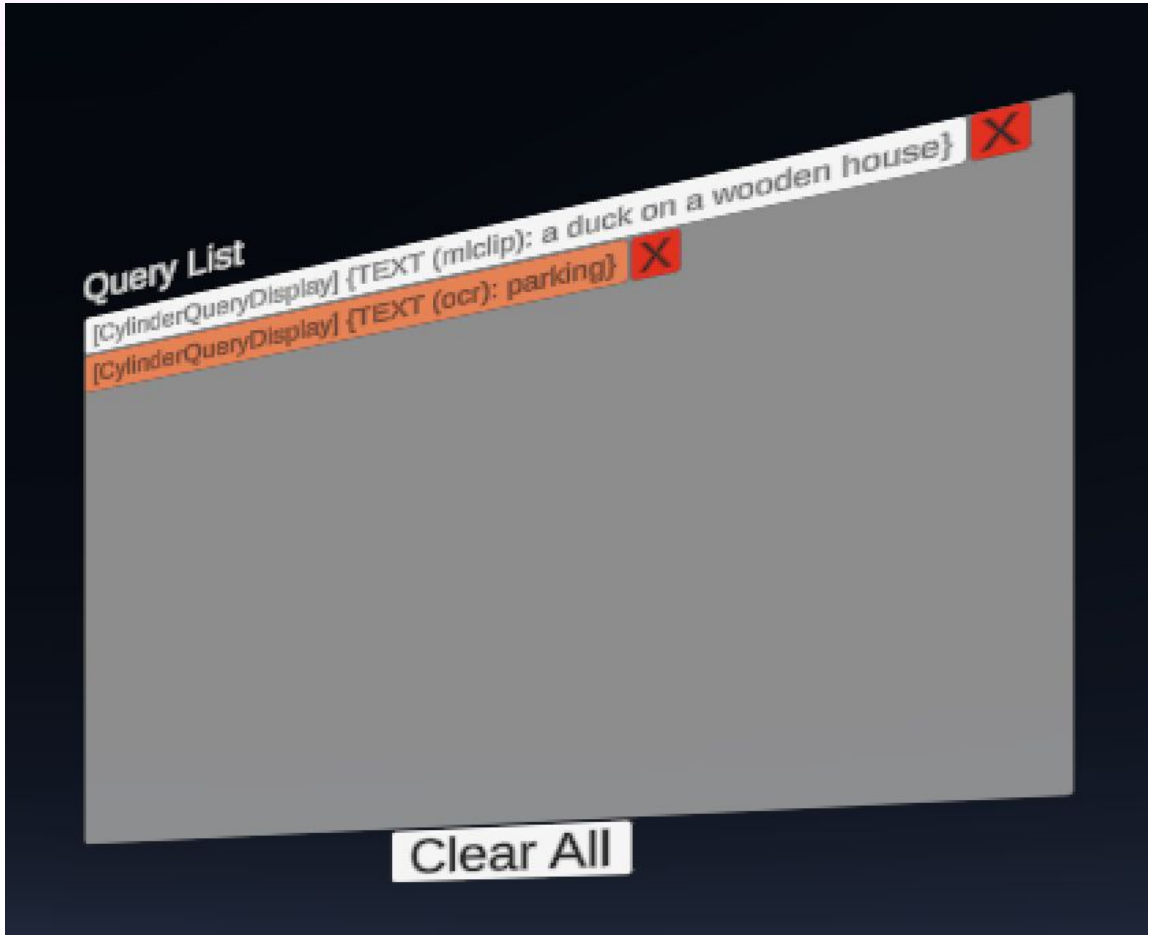


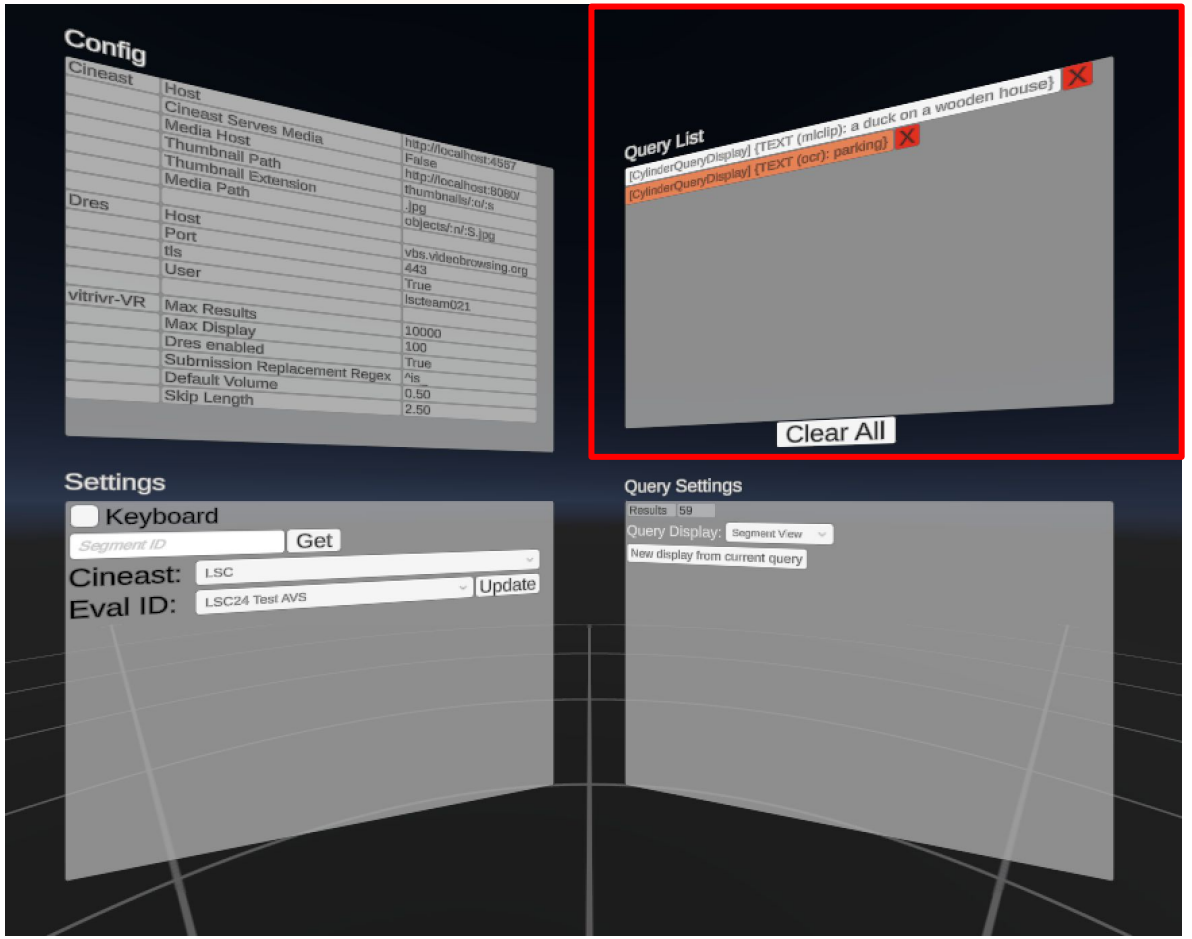
Options Menu: Configuration Information



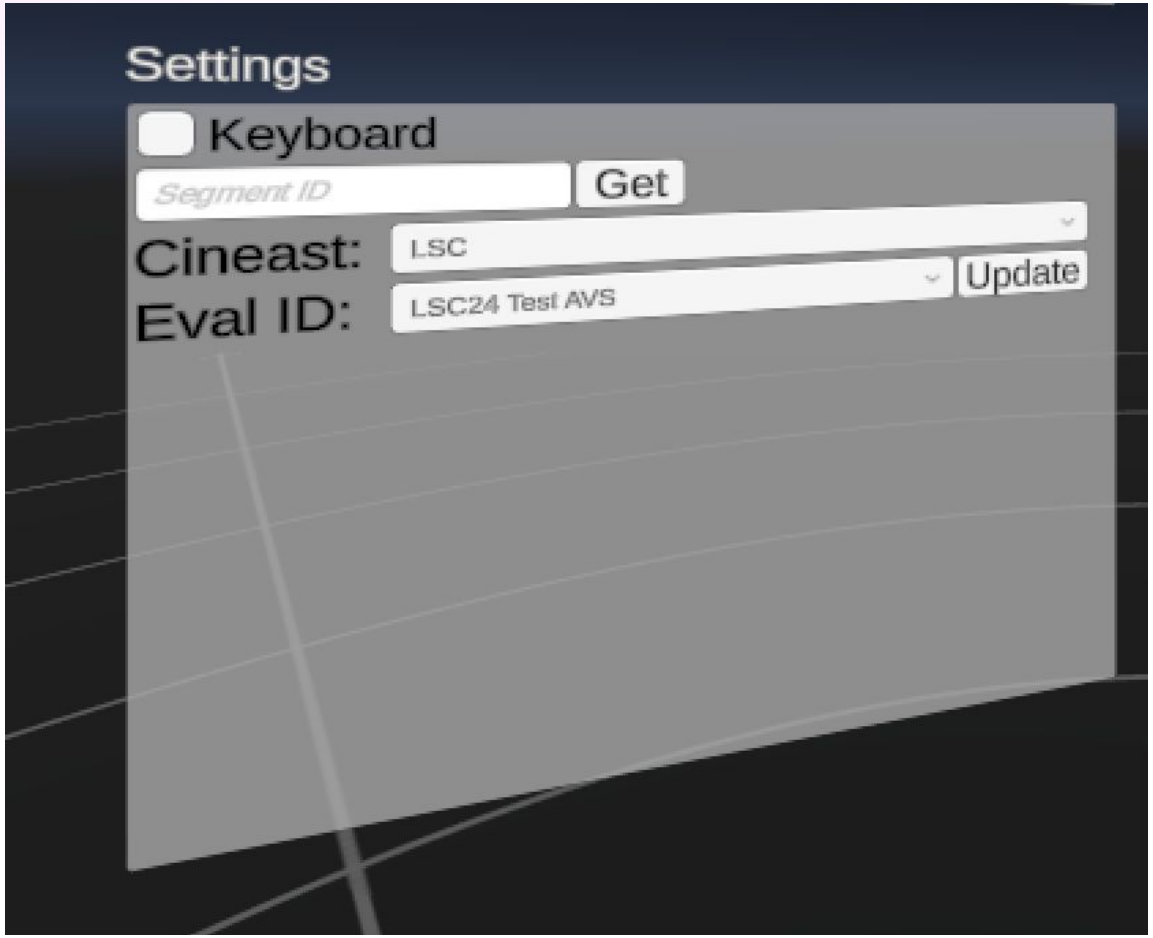


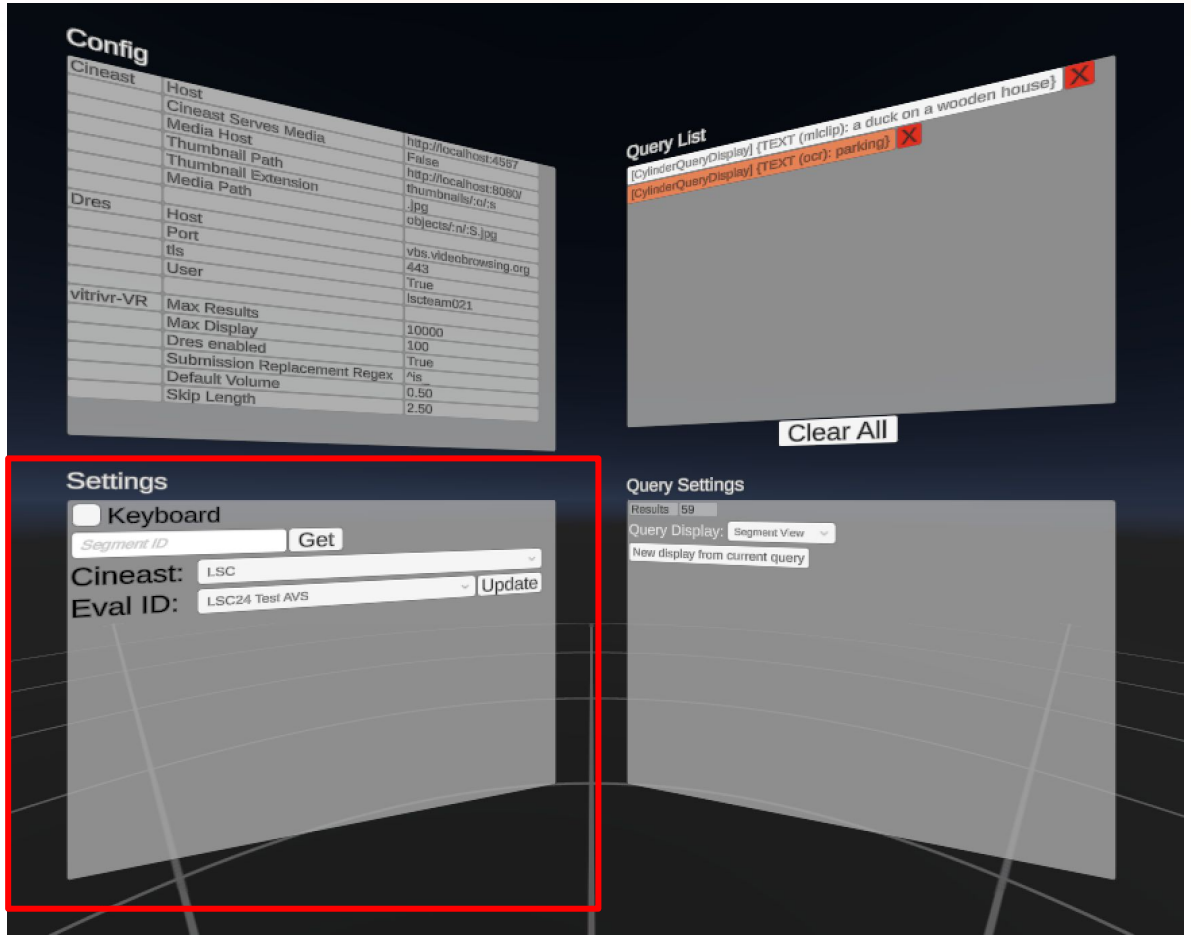
Options Menu: Query List



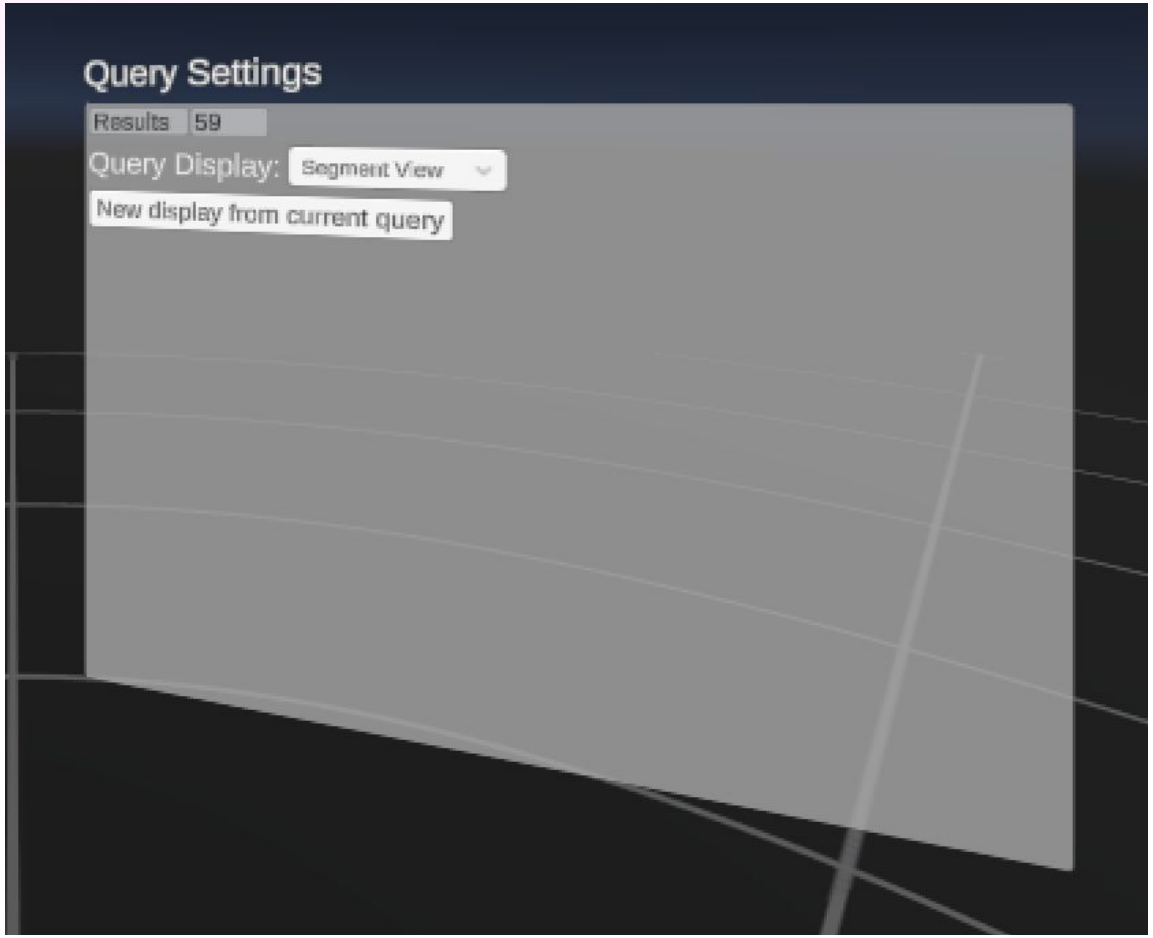


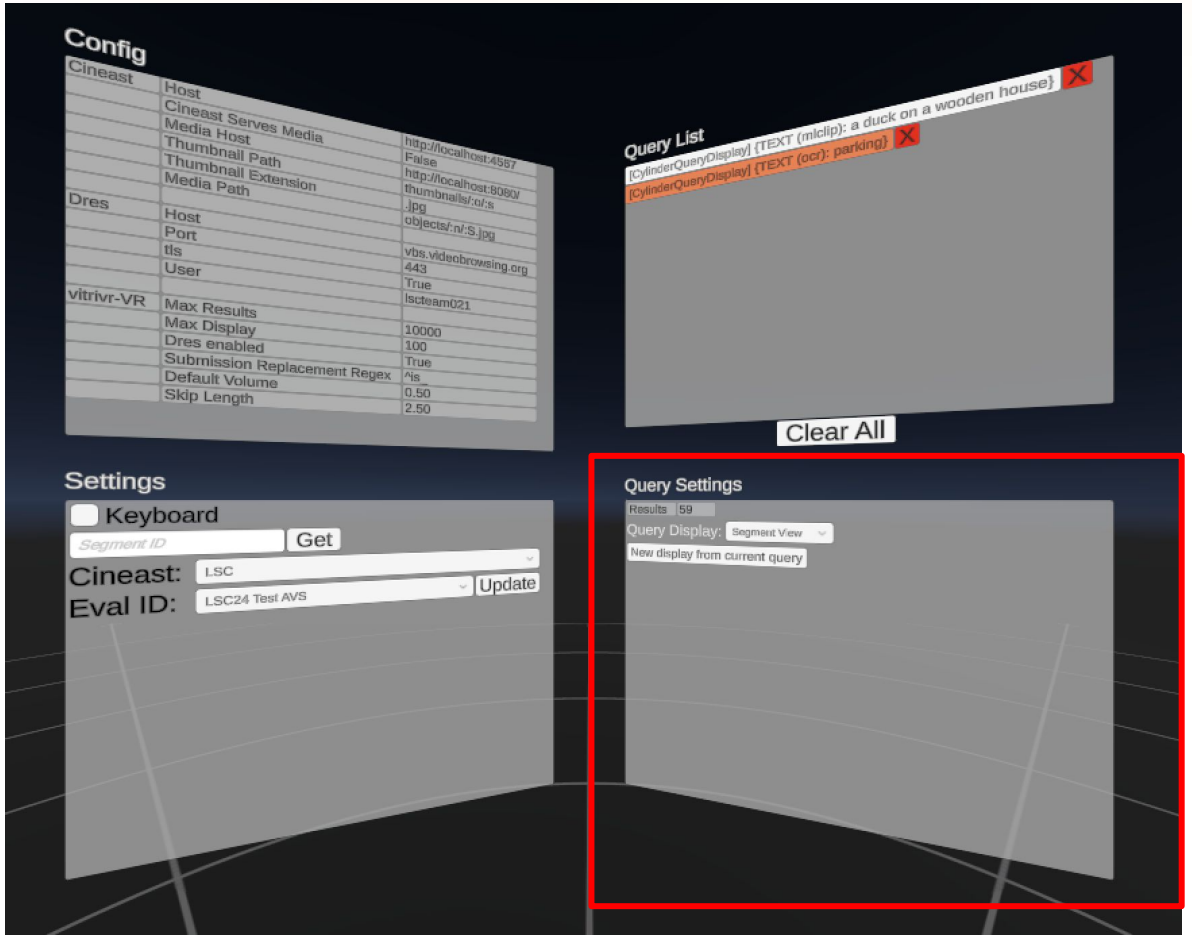
Options Menu: General Settings





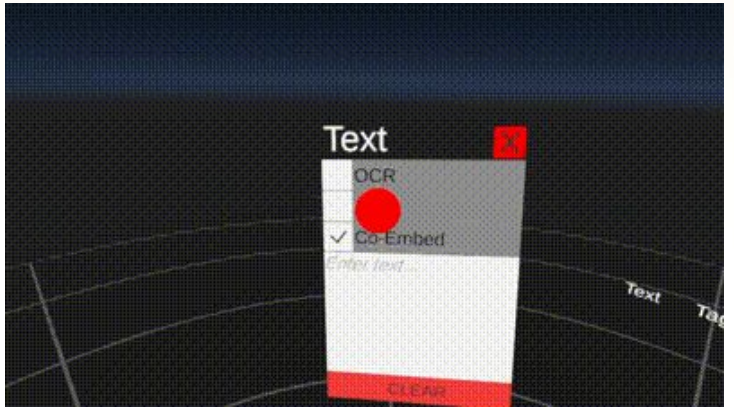
Options Menu: Query Settings

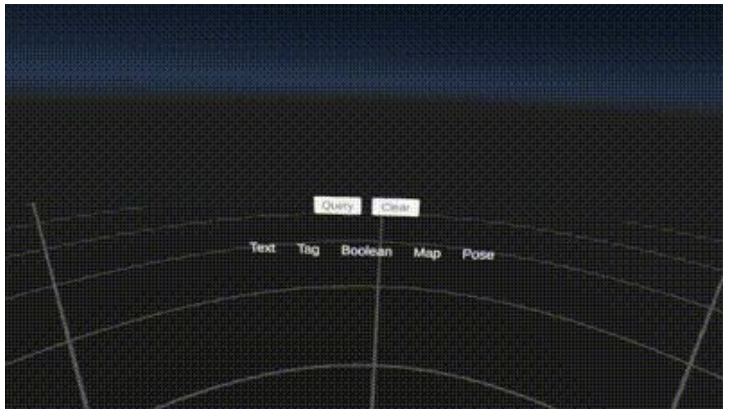


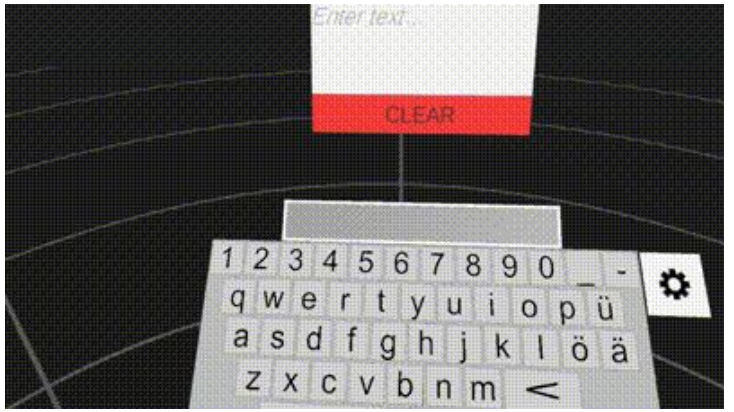


Exploring Query Formulation

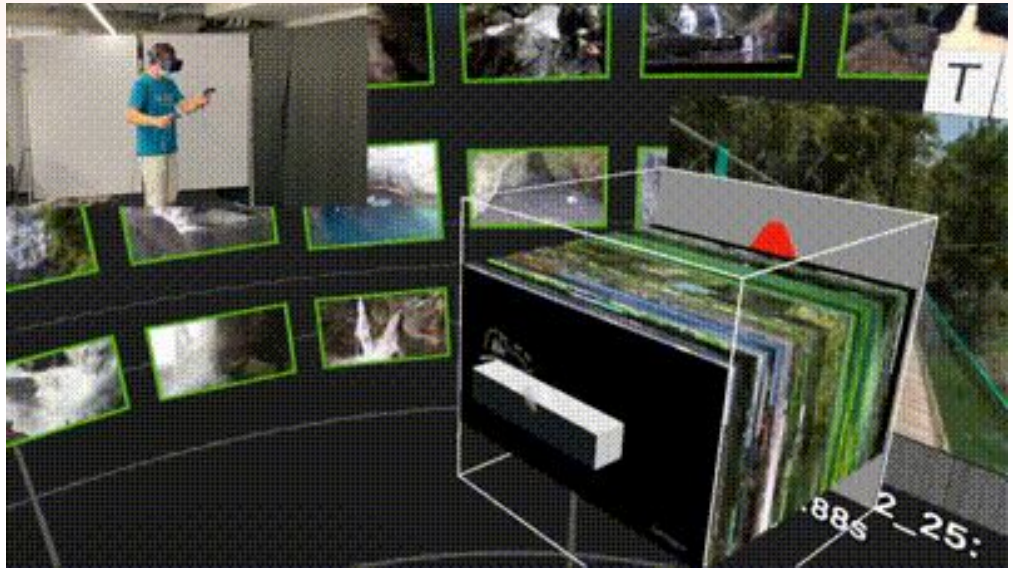
- Multimodal query formulation
  - Text (Embedding, OCR)
  - Boolean
  - Geospatial
  - Pose queries

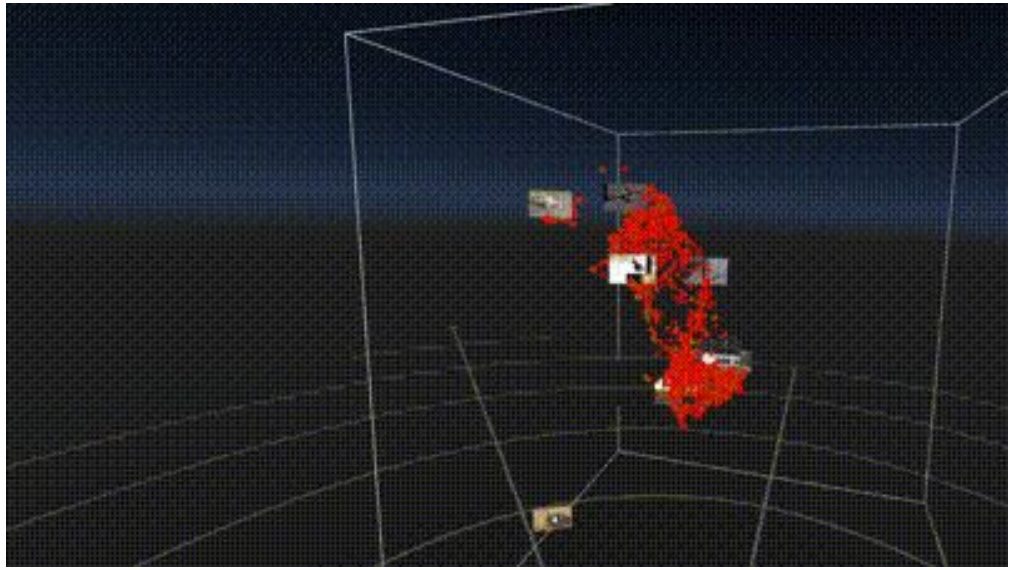






Advanced Results Exploration





References (1/3)

[A21] Ayush K. et al. “Geography-Aware Self-Supervised Learning,” ICCV 2021
[A90] Arnold, Stephen E. "The large data construct: a new frontier in database design." Microcomputers for Information Management 7.3 (1990): 185-203.
[B20] Börlin, S., Gasser, R., Spiess, F., Schuldt, H. (2020). 3D Model Retrieval Using Constructive Solid Geometry in Virtual Reality. AIVR 2020
[B22] Berton G. et al. “Deep Visual Geo-localization Benchmark,” CVPR 2022
[B95] Benford, Steve, et al. "VR-VIBE: A virtual environment for co-operative information retrieval." Computer Graphics Forum. Vol. 14. No. 3. Edinburgh, UK, 1995.
[C11] Chen, D. M. et al. “City-scale landmark identification on mobile devices“ CVPR 2011.
[C24] Cheng, T., Song, L., Ge, Y., Liu, W., Wang, X., Shan, Y.: YOLO-World: Real-Time Open-Vocabulary Object Detection. CoRR abs/2401.17270 (2024)
[D18] Duan, Ling-Yu, et al. "Compact descriptors for video analysis: The emerging MPEG standard." IEEE MultiMedia 2018.
[F93] Fitzmaurice, George W., Shumin Zhai, and Mark H. Chignell. "Virtual reality for palmtop computers." ACM Transactions on Information Systems (TOIS) 11.3 (1993): 197-218.
[Feng19] Feng, Y., Feng, Y., You, H., Zhao, X., & Gao, Y. (2019). MeshNet: Mesh neural network for 3D shape representation. In Proceedings of the AAAI Conference on Artificial Intelligence (AAAI) (Vol. 33, No. 01, pp. 8279-8286).
[Gezawa20] Gezawa, A. S., Zhang, Y., Wang, Q., & Yunqi, L. (2020). A review on deep learning approaches for 3D data representations in retrieval and classifications. IEEE access, 8, 57566-57593.
[G18] Giangreco, I. Database Support for Large-Scale Multimedia Retrieval, PhD Thesis, University of Basel, 2018.
[G20] Gasser, R., Rossetto, L., Heller, S., Schuldt, H. (2020). Cottontail DB: An Open Source Database System for Multimedia Retrieval and Analysis. ACM Multimedia 2020
[ILR] https://ilr-workshop.github.io/ECCVW2022/
[Jing21] Jing, L., Vahdani, E., Tan, J., & Tian, Y. (2021). Cross-modal center loss for 3D cross-modal retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 3142-3151).
[K21] Kordopatis-Zilos G. et al. “Leveraging EfficientNet and Contrastive Learning for Accurate Global-scale Location Estimation,” ICMR 2021.
[K23] Kerbl, B., et al. "3d gaussian splatting for real-time radiance field rendering." ACM Transactions on Graphics 42.4 (2023): 1-14.
[L21] Liu, Z., et al. “Swin transformer: Hierarchical vision transformer using shifted windows,“ CVPR 2021.

References (2/3)

[L22] J. Lokoč et al: A Task Category Space for User-Centric Comparative Multimedia Search Evaluations, MMM 2022
[LR18] https://landmarksworkshop.github.io/CVPRW2018/
[LR19] https://landmarksworkshop.github.io/CVPRW2019
[LR20] https://ilr-workshop.github.io/ECCVW2020/
[LR21] https://ilr-workshop.github.io/ICCVW2021/[M20] Mildenhall, Ben, et al. "NeRF: Representing Scenes as Neural Radiance Fields for View Synthesis." ECCV, 2020.
[ME] http://multimediaeval.org/
[N17] Noh, H. et al. “Large-scale image retrieval with attentive deep local features, “ CVPR 2017.
[N17] Noh, H. et al. “Large-scale image retrieval with attentive deep local features, “ CVPR 2017.
[N24a] Neuschmied, H., and W. Bailer. "Mining Landmark Images for Scene Reconstruction from Weakly Annotated Video Collections," MMM 2024.
[N24b] Neuschmied, H., and W. Bailer. "Efficient Few-Shot Incremental Training for Landmark Recognition”, Video4IMX @ ACM IMX 2024.
[P07] Philbin, J., et al., “Object retrieval with large vocabularies and fast spatial matching, “ CVPR 2007.
[P08] Philbin, J. et al. “Lost in quantization: Improving particular object retrieval in large scale image databases,” CVPR 2008.
[Pegia24] Pegia, M., Jónsson, B. Þ., Moumtzidou, A., Diplaris, S., Gialampoukidis, I., Vrochidis, S., & Kompatsiaris, I. (2024). Multimodal 3D Object Retrieval. In International Conference on Multimedia Modeling (MMM) (pp. 188-201). Cham: Springer Nature Switzerland.
[Qi21] Qi, S., Ning, X., Yang, G., Zhang, L., Long, P., Cai, W., & Li, W. (2021). Review of multi-view 3D object recognition methods based on deep learning. Displays, 69, 102053.
[R16] Rossetto, L., Giangreco, I., Heller, S. Tanase, C., Schuldt, H. (2016). Searching in Video Collections Using Sketches and Sample Images - The Cineast System. MMM (2) 2016
[R18] Rossetto, L. Multi-Modal Video Retrieval, PhD Thesis, University of Basel, 2018.
[R23] Razali, M.N.B. et al, “Landmark recognition model for smart tourism using lightweight deep learning and linear discriminant analysis, “ IJACSA 2023.

References (3/3)

[S23a] Spiess, F., Heller, S., Rossetto, L., Sauter, L., Weber, P., Schuldt, H. Traceable Asynchronous Workflows in Video Retrieval with vitrivr-VR. MMM (1) 2023.

[S23b] Spiess, F., Gasser, R., Heller, S., Schuldt, H. Rossetto, L. (2023). A Comparison of Video Browsing Performance between Desktop and Virtual Reality Interfaces. ICMR 2023

[SF16] Schonberger, Johannes L., and Jan-Michael Frahm. "Structure-from-motion revisited." ICPR. 2016.

[T13] Torii, A. et al. “Visual place recognition with repetitive structures,“ CVPR 2013.

[T15] Torii, A. et al. “24/7 place recognition by view synthesis,“ CVPR 2015.

[Visu] https://visulise.com/how-long-does-it-take-to-make-a-3d-model/

[W20] Wang, Xin, et al. "Frustratingly simple few-shot object detection." Proceedings of the 37th International Conference on Machine Learning. 2020.

[W20a] Weyand, T. et al. “Google landmarks dataset v2-a largescale benchmark for instance-level recognition and retrieval, “ CVPR 2020.

[W20b] Warburg F. et al. “Mapillary street-level sequences: A dataset for lifelong place recognition” CVPR 2020

[Y18] Yu, F, “Bdd100k: A large-scale diverse driving video database,“ https://bair.berkeley.edu/blog/2018/05/30/bdd (2018).

[Y22] Yang, M. et al. “2nd place solution to google landmark retrieval 2020,” arXiv:2210.01624, 2022.

[Z10] Zamir, A. R., “Accurate image localization based on google maps street view, “ ECCV 2010.

[Z20] Zhuang, Peiqin, et al. "Learning attentive pairwise interaction for fine-grained classification." AAAI 2020.

Thank you!

XReco.eu

Materials of the tutorial:

https://github.com/XRecoEU/MultimediaRetrievalXR

Co-financed by the EC under Grant Agreement ID: 101070250