# Metaverse Retrieval:
# Finding the Best Metaverse Environment via Language

Ali Abdari
abdari.ali@spes.uniud.it
University of Udine
Udine, Italy
University of Naples Federico II
Naples, Italy

Alex Falcon
falcon.alex@spes.uniud.it
University of Udine
Udine, Italy

Giuseppe Serra
giuseppe.serra@uniud.it
University of Udine
Udine, Italy

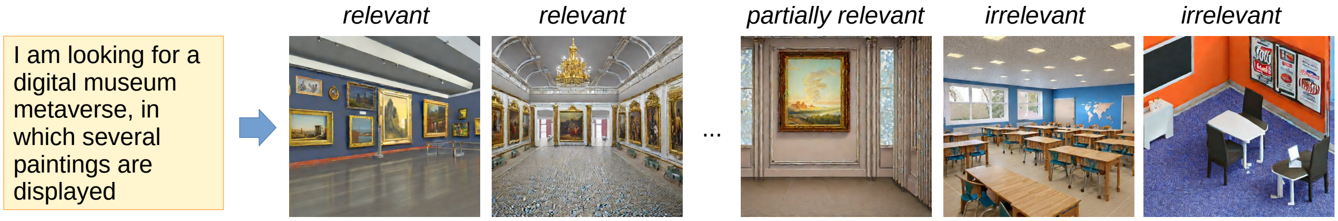
I am looking for a digital museum metaverse, in which several paintings are displayed
relevant
relevant
...
partially relevant
irrelevant
irrelevant

**Figure 1: Overview of the proposed text-to-metaverse retrieval problem: given a textual query, the output is a ranking list of metaverse scenarios arranged in descending order according to relevance. Details in Section 3.**

## ABSTRACT

In recent years, the metaverse has sparked an increasing interest across the globe and is projected to reach a market size of more than $1000B by 2030. This is due to its many potential applications in highly heterogeneous fields, such as entertainment and multimedia consumption, training, and industry. This new technology raises many research challenges since, as opposed to the more traditional scene understanding, metaverse scenarios contain additional multimedia content, such as movies in virtual cinemas and operas in digital theaters, which greatly influence the relevance of the metaverse to a user query. For instance, if a user is looking for Impressionist exhibitions in a virtual museum, only the museums that showcase exhibitions featuring various Impressionist painters should be considered relevant. In this paper, we introduce the novel problem of text-to-metaverse retrieval, which proposes the challenging objective of ranking a list of metaverse scenarios based on a given textual query. To the best of our knowledge, this represents the first step towards understanding and automating cross-modal tasks dealing with metaverses. Since no public datasets contain these important multimedia contents inside the scenes, we also collect and annotate a dataset which serves as a proof-of-concept for the problem. To establish the foundation for it, we implement and analyze several solutions based on deep learning, whereas to promote transparency and reproducibility, we will publicly release their source code and the collected data.

*MMIR '23, November 2, 2023, Ottawa, ON, Canada*
© 2023 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0271-6/23/11.
https://doi.org/10.1145/3606040.3617445

## CCS CONCEPTS

• **Information systems** → **Multimedia and multimodal retrieval**.

## KEYWORDS

Metaverse, Scene Retrieval, Multimedia Retrieval, Contrastive Learning

**ACM Reference Format:**
Ali Abdari, Alex Falcon, and Giuseppe Serra. 2023. Metaverse Retrieval: Finding the Best Metaverse Environment via Language. In *Proceedings of the 1st International Workshop on Deep Multimodal Learning for Information Retrieval (MMIR '23), November 2, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3606040.3617445

## 1 INTRODUCTION

The term 'metaverse' refers to a hypothetical virtual universe blending synthetic and digital elements with the real world [36]. To do so, it is facilitated by Web technologies and 'extended reality', an umbrella term encompassing virtual, augmented, and mixed reality technologies [25, 36]. The rapidly developing metaverse market, is currently valued at $65B, and is projected to exceed $1000B by 2030 [6]. As a result, the metaverse represents a new frontier of digital experiences, providing immersive environments with a multitude of applications, including education, industry, and entertainment [11, 47]. For example, users can socialize in engaging environments using applications like VRChat[1] and Meta Horizon World[2], or doctors can simulate complex surgical procedures in fail-proof environments like MetaMedicsVR[3]. In particular, we are seeing a significant growth in metaverses related to entertainment and art, which are increasingly used as platforms to explore our shared history. These

[1] https://hello.vrchat.com/
[2] https://www.meta.com/it/horizon-worlds/
[3] https://metamedicsvr.com/

include historical site visits (e.g., VersaillesVR[4], The Anne Frank House VR[5]), digital museum tours (e.g., Musée Dezentral[6], VOMA[7], Museum of Crypto Art[8]), and even creating personalized art in virtual environments (e.g., Painting VR[9], Vermillion[10]). The emergence of new metaverses necessitates advanced retrieval methods, similar to the trend that led to Google Images and YouTube search engines for user-generated content.

Unlike other forms of media such as images, videos, and 2D and 3D scenes, metaverse scenarios present a more complex and challenging environment due to their richness in multimedia elements, and their dynamic nature. These characteristics introduce new challenges and necessitate understanding from both the scene and multimedia domains. For instance, distinguishing between a car showroom metaverse and a digital museum may be accomplished by reasoning on scenario-level features, i.e., characteristics extracted from scene understanding methods, though finer-grained details (e.g., the absence of paintings in the former) also serve as discriminators. Conversely, scenario-level features may not suffice when comparing an Impressionist exhibition and a modern art museum, thereby significantly elevating the importance of the multimedia content present in the metaverse. Moreover, these contents play a crucial role in relation to the dynamic nature of metaverses: exhibitions at digital museums, for example, would often change over time, and their relevance to user interests would therefore vary. In light of these emerging challenges, in this paper, we define the novel problem of retrieving metaverses based on user-defined textual queries and introduce the task of text-to-metaverse retrieval. Similar to other cross-modal retrieval tasks, this task requires ranking a list of metaverses based on their semantic relevance to an input textual query, as can be seen in Figure 1, yet assessing such relevance is greatly influenced by the multimedia content present in the scenarios. Additionally, since existing datasets are unsuitable for the task due to their lack of multimedia content, we also collected a dataset of over 3000 multimedia-enriched scenarios, each paired with a textual description, to establish a proof-of-concept and a benchmarking opportunity for the text-to-metaverse retrieval task. Furthermore, we develop a deep learning-based framework for this task and investigate different models, assessing and comparing their performance.

Therefore, our major contributions can be summarized as follows:

- We define the problem of retrieving metaverse scenarios by means of a natural language description, which requires an understanding of both the scene and the multimedia contents shown in it. It is a fundamental task to enhance and support the search process done by the users to find a metaverse which satisfies their needs;
- Since the scenes in previous public datasets do not contain any multimedia content, we collect a dataset for benchmarking purposes of this novel task, containing more than 3000 multimedia-enriched scenarios, and annotate each of them with a textual description describing the furniture and their positional relations, and the multimedia contents showcased in it. The dataset is available on GitHub to promote research on this topic;
- We design and develop a deep learning-based framework to align the scene and multimedia features to their textual counterpart. To support the reproducibility of the results, we publicly released all the code on GitHub.

In Section 2 we describe previous research areas which are related to the metaverse and cross-modal retrieval. Section 3 introduces the novel text-to-metaverse retrieval task and describes the dataset we created, whereas the implemented methods are detailed in Section 4. Several experimental results are proposed and discussed in Section 5. Finally, Section 6 concludes the manuscript and proposes future research directions.

## 2 RELATED WORKS

The research work related to the novel task of text-to-metaverse retrieval can be divided into three macro areas. The first one is focused on the recent advancements in metaverse-related research. In the second one, we discuss the relations with the problem of scene understanding although, as mentioned in the introduction, metaverse scenarios pose a more challenging environment. In the third one, we explore the relations of the proposed task to the cross-modal retrieval task: in fact, since we aim to retrieve metaverses by means of a natural language query, it is fundamental to model the underlying interactions between the two modalities.

### 2.1 Research on metaverse

As metaverses become more immersive and consumer-level extended reality technologies become available, research on potential applications and use cases is increasing. Many of the popular applications are related to virtual try-on and shopping [15, 54], digital museums [12, 35], education [5, 17], and training, e.g., in surgery [33] and industrial maintenance [53]. Moreover, even more complex applications could be implemented by relying on digital twins, which are digitalization of physical entities to which they are still bonded, allowing for reciprocal influence in case any change is performed on either of them [23]. For instance, applications related to machine monitoring and fault prediction [4, 20] and smart healthcare of elder people [34, 40]. Noteworthily, all these applications are supported by the advancements made in recent years in computer vision, achieving more effective human pose estimation [52, 58], semantic segmentation [9, 26], and object detection [37, 43].

However, to the best of our knowledge, no approach has been developed to filter the metaverses based on their attributes, both in terms of environment and multimedia content, and their suitability for a given user query, so the users must perform the search by themselves. To address this limitation, in this paper, we introduce and address the task of text-to-metaverse retrieval.

### 2.2 Scene Understanding

Scene understanding aims at enabling machines to comprehend and interact seamlessly with the real world, which has resulted in considerable research interest due to its challenging environment

[4] https://en.chateauversailles.fr/news/life-estate/versaillesvr-palace-yours
[5] https://annefrankhousevr.com/
[6] https://musee-dezentral.com/
[7] https://www.voma.space/
[8] https://museumofcryptoart.com/
[9] https://www.paintingvr.xyz/
[10] https://vermillion-vr.com/

and the direct impact on industry, and societal assistance, e.g., related to human-robots cooperation, autonomous driving cars, etc. Computers need to be able to analyze the structure and layout of a scene by processing diverse types of information from different sources, such as multi-view images, 3D meshes, or point clouds, in order to gain a deep understanding of a three-dimensional setting. To address such a challenging problem, several sub-tasks which act as building blocks for more complex applications were identified, such as object detection [18, 60], segmentation [22, 56], depth estimation [48], and semantic understanding [27]. These important advancements in scene understanding have a considerable impact on real-world applications, such as autonomous driving [39] and robotics planning [24]. Moreover, they were shown to be effective in applications related to augmented reality [32] and virtual reality [16], allowing for their use also in metaverse-related applications.

However, scenes and metaverses constitute different data types. The former involves an environment comprising objects and furniture, whereas the latter is an immersive scenario incorporating a variety of multimedia content, such as visual artworks, and TV programs. Due to this fundamental difference, in this paper, we both introduce a novel and more challenging task and also collect a proof-of-concept dataset, as existing datasets are unsuitable due to their lack of multimedia content in the scenes. This dataset contains multimedia-enriched scenarios along with detailed descriptions of their furniture and the multimedia content within each scenario.

## 2.3 Cross-modal retrieval

Given the enormous amounts of content (e.g., videos, images, but also metaverses) created and uploaded daily to the Internet, it is fundamental to be able to retrieve those relevant to the users' interests and filter out the irrelevant content. For instance, more than 500 hours of video are uploaded to YouTube every minute [7], and 95 million videos and images are uploaded to Instagram daily [44]. To inform the search engine about what contents are relevant, including metaverses, users formulate their needs by means of a natural language query, therefore requiring cross-modal retrieval, i.e., a search process across different data modalities. In fact, cross-modal retrieval enables this task by means of techniques using one modality, typically text, to query a search engine and rank the elements of the other modality, such as image, video, and audio, based on their relevance to the query. Deep learning techniques are often used to automatically discover complex relationships between the inputs obtaining highly performing cross-modal retrieval methods, e.g. when using text to retrieve videos [19, 42], images [10, 50], and audio [31, 41]. These methods learn how to map the multimodal inputs into a joint embedding space in which the representations of paired inputs, e.g., images and its own textual description, are similar, i.e., close in the embedding space, resulting in efficient retrieval via cosine similarity or other ranking functions. To learn the joint embedding space, contrastive loss functions are used since they aim at increasing the similarity for paired inputs while decreasing it for unpaired ones [45, 51]. Inspired by these works, in this paper, we introduce the novel task of text-to-metaverse retrieval and implement a cross-modal retrieval method by building a joint text-metaverse embedding space. Noteworthily, it is a novel problem since, to the best of our knowledge, previous works on scene retrieval either located the objects in the scene [46], the text shown inside them [57, 61], or retrieved 3D scenes using 2D images [2, 3] or sketches [65]. The closest work to our contribution are the SHREC 2018/2019 challenges [2, 3], yet there are two main differences: first, in SHREC the queries are represented by images, making the approaches developed for it are less flexible and not directly usable in our setting, as they require the user to own a picture of the desired metaverse; second, the scenes in SHREC do not feature any multimedia content, making them unsuitable to represent metaverses. These limitations in existing literature further motivate the collection of our dataset.

# 3 PROPOSED TASK: TEXT-TO-METAVERSE RETRIEVAL

In this work, we propose a novel task which we name text-to-metaverse retrieval, taking inspiration from recent advancements in multimedia-related fields, which can be described as follows. The data considered for this task is made of metaverse-description pairs, $\mathcal{D} = \{(m_1, d_1), \dots, (m_N, d_N)\}$, in which textual paragraphs describe each of the metaverse scenarios by highlighting the multimedia contents in it, e.g., static (paintings) and dynamic objects (movies), and the many pieces of furniture which embellish the environment (e.g., tables, beds, and TVs). The task objective is to produce a ranking list of the metaverses based on their relevance to the description which is used as a query: this means that, given the description $d_i$ the resulting ranking list will be a permutation of the metaverse scenarios, $\pi = m_{\pi_1}, \dots, m_{\pi_N}$, which should have the corresponding metaverse, $m_i$, at its top rank, i.e., $m_{\pi_1} = m_i$. Nonetheless, it is interesting to note that multiple metaverses may be at least partially described by the same query and that several descriptions may be relevant for the same metaverse. Therefore, for evaluation purposes it is recommended to use two types of metrics: first, the recall rates and median rank, as is typically done for other multimedia retrieval tasks, e.g., in text-to-image [38, 50] and text-to-video retrieval [19, 59], which are used to quantify the capabilities of a model to retrieve the correct ground truth element; second, more complex metrics inspired from the information retrieval field, such as the Normalized Discounted Cumulative Gain (nDCG), is considered since, as recently explained for text-to-video retrieval [62], it may capture more complex behaviors related to a higher level understanding of the semantics of the data. Furthermore, the opposite task, i.e., metaverse-to-text retrieval may also be considered to obtain a holistic evaluation of the model performance.

## 3.1 Dataset collection

To address the novel text-to-metaverse retrieval task, we gathered a large-scale set of metaverse-description pairs. Considering that metaverses can be seen as scenes with additional multimedia content in them, we looked for suitable datasets in the previously published literature on scene-related topics. However, none of them featured the multimedia aspect, making them unsuitable for the proposed task and therefore motivating the collection of our dataset.

*3.1.1 Metaverse scenarios.* Inspired by the amount of metaverses which recreate the houses of important people of the past, we decided to reuse already available data by gathering 3384 professionally designed indoor scenarios from the 3D-front dataset [21],

**Figure 2: The floor plans of some of the sample scenes**

thereby including a diverse range of furniture and object types commonly found in indoor environments, such as beds, wardrobes, tables, cabinets, and TV-stands. However, as mentioned before, these scenes do not feature multimedia content. Therefore, to simulate the presence of multimedia content in these metaverse scenarios, we included a video along with a TV placed on suitable objects such as "TV stands" or "tables". Specifically, we picked the visual content from a subset of 25 randomly-picked YouCook2 videos [66] and uniformly distributed them into the metaverses under analysis. We chose to use videos as a form of multimedia content because recreated houses from the past often feature educational videos to inform the visitors about the previous dwellers and their life, habits, and historical context. In the 2 the floor plans of some of the 3d scenes used in this work can be seen.

*3.1.2 Textual descriptions.* To provide a textual description for each of the metaverses, we use the category (e.g., "Table", "Sofa", etc), style (e.g., "Japanese", "Modern", etc), theme (e.g., "Lines", "Smooth Net", etc), and material (e.g., "Wood", "Leather", etc) annotations provided with the 3D-front dataset. To do so, we first prepare three sets of sentences, $\mathcal{O}$, $\mathcal{P}$, and $\mathcal{V}$, automatically. In the first set, each metaverse scenario is described by several sentences capturing the existing objects' features separately. For instance, a sentence in this set may look like this: "This room contains two Composite Smooth Net Japanese Dining Tables". In the second set of sentences, $\mathcal{P}$, we capture the distance between existing objects in each scenario, indicating how far or how close two objects are in a scene based on the standard deviation of the distribution of distances across all scenes. For instance, "Solid Wood Japanese Nightstand is close to the Smooth Leather Lines Modern Multi-seat Sofa". In the final set, $\mathcal{V}$, the visual contents of the video are described by means of the captions present in the video dataset and further preprocessed. In particular, the spaCy library [1] was used to make the captions more harmonious with the previous sentences: to do so, we detect the verbs by using spaCy's part-of-speech tagger and then adapt them to the context. This means that after this additional preprocessing, for example from the caption "cut jalapeno peppers and remove the seeds" we obtain the sentence "This room contains a TV showing cutting jalapeno peppers and removing the seeds". Ultimately, we obtain the final description as a concatenation of the three, that is $d_i = [o_i, v_i, p_i]$, resulting in a list of sentences per metaverse whose length ranges from 6 to 68 (17.5 on average). An overview of the statistics of the textual annotations can be seen in Figure 3.

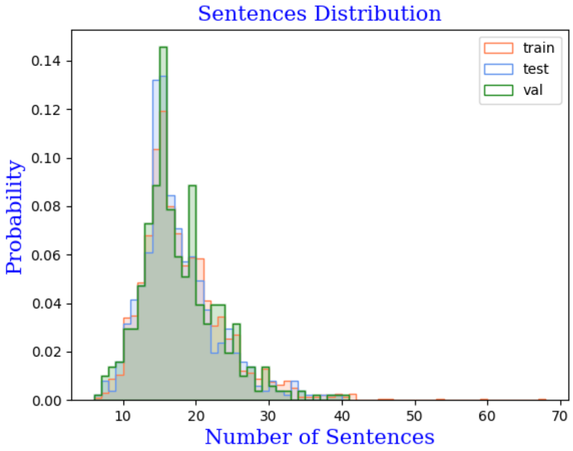

**Figure 3: Distribution of the number of sentences per scene in the train, validation, and test splits.**

## 4 PROPOSED METHOD

To address the text-to-metaverse retrieval task, we propose a method which uses deep learning to first extract a representation both for the metaverse and the associated description, and then to increase their similarity in a cross-modal embedding space. By doing so, it will be possible to search and rank the scenes by mapping the query into the same embedding space using the functions learned at training time. An overview of the proposed method is shown in Figure 4. As mentioned in Section 3, we consider a dataset $\mathcal{D}$ composed of metaverse scenarios and associated descriptions (the details can be found in Section 3.1).

Considering that metaverses can be seen as a scene with additional multimedia content, we decided to first model these two sources of information separately and then fuse the acquired knowledge. To model the metaverse scenario, we use an Auto Encoder-based approach based on recent advancements in scene representation learning via deep Variational Auto Encoder (VAE) [64]. In fact, in recent years deep Auto Encoders were often used to automatically learn compact yet highly discriminative representations for scenes [30, 64] and also other types of data, including video [8, 63] and audio [14, 49]. In particular, we use a pre-trained VAE [64] to obtain the representation $\phi_S$ for the scenario, whereas the representation $\phi_V$ for the video is obtained through an Auto Encoder, which is learned on top of the visual features extracted from a pre-trained deep neural network, S3D [45]. The autoencoder is defined as follows:

$$\begin{aligned}\phi_{v_1} &= tanh(W_{v_1}\phi_{v_0} + b_{v_1})\\ \phi_V &= tanh(W_{v_2}\phi_{v_1} + b_{v_2})\end{aligned}$$

where $\phi_{v_0}$ is the deep representation for the video, and $W_*, b_*$ are trainable weights and biases.

The two representations are then concatenated, resulting in $\phi_M$, and combined together by using a fully-connected network, named FCNet, as follows:

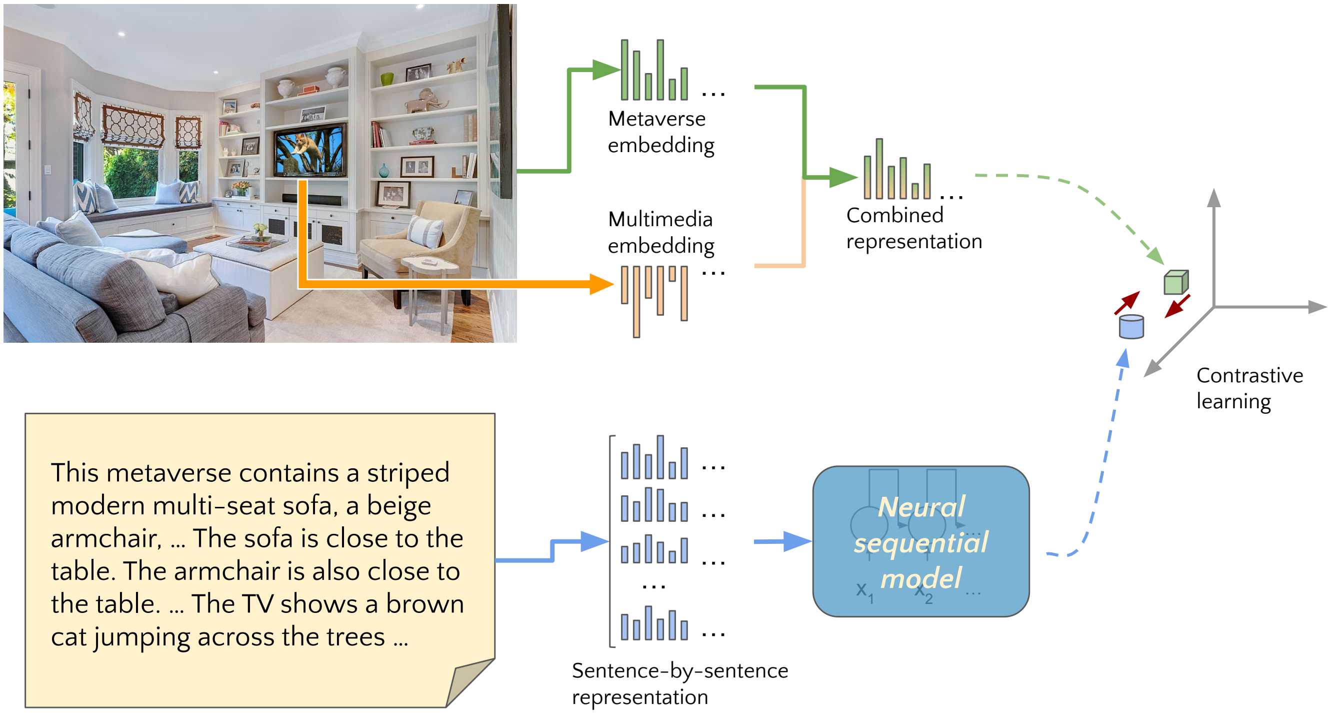
Metaverse embedding
Multimedia embedding
Combined representation
Contrastive learning
This metaverse contains a striped modern multi-seat sofa, a beige armchair, ... The sofa is close to the table. The armchair is also close to the table. ... The TV shows a brown cat jumping across the trees ...
Neural sequential model
Sentence-by-sentence representation

**Figure 4: Overview of the proposed method. Details in Section 4.**

$$\phi_{m_1} = BN(\delta_1(ReLU(W_{f_1}\phi_M + b_{f_1})))$$
$$\phi_{m_2} = BN(\delta_2(ReLU(W_{f_2}\phi_{m_1} + b_{f_2})))$$
$$\rho_M = W_{f_3}\phi_{m_2} + b_{f_3}$$

where $W_*, b_*$ are trainable weights and biases, $\delta_*(\cdot)$ represents the dropout operator, $BN(\cdot)$ identifies the use of Batch Normalization [28], and $\rho_M \in \mathbb{R}^{1\times D}$ is the descriptor obtained for the metaverse, which is then used to perform the learning.

Given that a metaverse scenario may be highly complex and detailed, describing it may lead to a paragraph made of many sentences, e.g., more than 50 sentences as in our dataset (see Section 3.1). Therefore, extracting discriminant descriptors from these paragraphs containing several thousands of tokens may become difficult even for highly performing methods which are often limited in the number of tokens they can process, e.g., 512 for BERT. To this end, we designed the following method, consisting of two major steps. First, we obtain a list of $M$ shorter sentences, each with a maximum of 512 tokens, by using the periods as a splitting point and extracting for each of them a sentence embedding using BERT, obtaining $\phi_T \in \mathbb{R}^{M\times 768}$. Second, a neural sequential model is used to capture temporal relations between them, resulting in a single descriptor, $\rho_Q \in \mathbb{R}^{1\times D}$. Specifically, to implement this model, several deep learning architectures were considered and thoroughly tested (see Section 5). First, we include a simple approach based on **mean pooling**, which we defined by the following operations:

$$\phi_{T_a} = \frac{1}{M}\sum_{i=1}^{M}\phi_T^{(i)}$$
$$\rho_Q = W_t\phi_{T_a} + b_t$$

in which we first compute the average of the sentences, $\phi_{T_a}$, followed by a trainable linear transformation. Then, considering that sequential patterns may lie hidden in the list of descriptions, Gated Recurrent Units (GRU) [13] represent a possible way to discover them: therefore, we both consider a **GRU** solution, in which the last hidden state, $\overrightarrow{h}$, is used as the final embedding, and a more sophisticated **bidirectional GRU**, in which the mean of $\overrightarrow{h}$ and $\overleftarrow{h}$ is taken. A final solution is based on Multi-head **Self-Attention** [55], which uses the attention mechanism in place of the recurrence used by the GRU to attend each sentence to a different degree. Then, the description's final embedding is computed by summing up the sentence representations weighted by their attention scores. This is done by using $\phi_T$ as the Q, K, and V in the following equations:

$$MultiHead(Q,K,V) = [\text{head}_1, \ldots, \text{head}_h]W^o$$
$$\text{head}_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$
$$Attention(q,k,v) = softmax(\frac{qk^T}{\sqrt{d_k}})v$$

where [] represents a concatenation, and $W_i^Q, W_i^K, W_i^V$ are trainable weights.

To perform cross-modal retrieval, a contrastive learning approach is often used to learn a joint embedding space in which the distance between corresponding feature vectors of metaverses and descriptions is minimized, while that of unrelated ones is maximized. By doing so, it is possible to form a new query by means of a natural language description, then mapped into the same embedding space, and used to rank the metaverses by using a similarity metric, such as the Euclidean distance. In particular, to train all of the models we used the triplet loss [51], defined as follows:

$$\mathcal{L}(a,p,n) = max(0, \Delta + s(a,n) - s(a,\ p))$$

$$\mathcal{L} = \frac{1}{2 \cdot N \cdot (N-1)} \sum_{i=1}^{N} \sum_{j=1, j\neq i}^{N} \mathcal{L}(\rho_{Q_i}, \rho_{M_i}, \rho_{M_j}) + \mathcal{L}(\rho_{M_i}, \rho_{Q_i}, \rho_{Q_j})$$

where $\Delta$ is a fixed margin, and $s(\cdot,\cdot)$ is the cosine similarity.

## 5 EXPERIMENTAL RESULTS

To examine the performance of the proposed methodology in text-to-metaverse retrieval task, we use the dataset described in Section 3.1. As mentioned before, it contains more than 3300 indoor scenarios containing different kinds of objects among twenty categories, like beds, wardrobes, tables, cabinets, and TV-stands, alongside the multimedia content which is shown in it.

In the training procedure, we used 70% of the data for the train set, while two 15% proportions were considered for validation and test sets. The main task addressed in this paper is to obtain a model capable of retrieving metaverses based on textual information, i.e., a text-to-metaverse model; however, to have a holistic evaluation of the performance, we also consider the metaverse-to-text task. As mentioned in Section 3, given a textual query (respectively, a metaverse) we consider two types of metrics. First, we use metrics which solely seek to measure the performance of the model in retrieving the groundtruth, including the recall at k, **R@k**, which aims at measuring how many times the groundtruth metaverse (resp., description) is located in the top k positions of the output ranking list, and the **Median Rank** of the groundtruth. Secondly, we also consider a metric which quantifies the quality of the output ranking list, that is we adopt the Normalized Discounted Cumulative Gain, **nDCG** [29], to estimate how close the ranking list is to the optimal one, which ranks the metaverses (resp., descriptions) following a descending relevance degree. When **nDCG@10** is reported, it means that the nDCG is only computed for the first 10 relevant metaverses (resp., descriptions) and not for all the relevant ones. Specifically, we represent both the metaverse and the query with a list of their objects, relations, and video. Then, if the metaverse contains at least half of the characteristics existing in the query, it is considered relevant to the query itself.

### 5.1 Implementation details

In our implementation, the dropout rates of $\delta_1$ and $\delta_2$ were empirically set to 0.2 and 0.15, respectively. The margin used to separate the groundtruth metaverse-description pair and the other elements in the embedding space was set to $\Delta = 0.25$. The optimizer used is Adam with default parameters. The learning rate was initially set to 0.008 and then reduced by 25% after 27 epochs. The training lasts 50 epochs, with a batch size of 64, after which we select the best-performing model on the validation set.

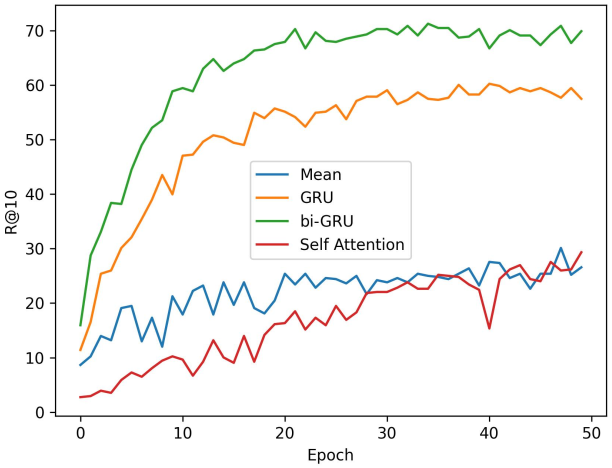

**Figure 5: Text-to-metaverse R@10 performance achieved on the validation set by each of the models under analysis. More details in Section 5.2.**

In experiments, we set $D = 400$ because by using the pre-trained VAE published by Yang et al. [64] we obtain a feature vector of size 200 and similarly we used the same size also for the representation of the video, can be obtained using of the autoencoder detailed in Section 4. In the Self-Attention model, we used $h = 16$ heads.

All the experiments were performed using a machine running an RTX A5000 GPU, 16 GB of RAM, and an Intel Xeon E5-1620. We used PyTorch 1.12.1 to implement the deep learning algorithms, whereas spaCy 3.1 was used for sentence processing. The source code and data are publicly available. [11]

### 5.2 Baselines comparison

As a first experiment, we compare the chosen baselines, in order to understand which model is best suited for the proposed task. On the validation set (Figure 5), it rapidly becomes clear that the bidirectional GRU model (green) achieves the best overall results, keeping a margin of around 9-10% in R@10 over the monodirectional GRU (orange). This result is also confirmed on the test set (Table 1), indicating both the effectiveness of the GRU-based model and of the bi-directionality, which helps in understanding the relations between the sentences forming the metaverse description. In fact, it can be seen that the bidirectional GRU model achieves 68.5% R@10 on the text-to-metaverse task, with a margin of around 10 points over the monodirectional one (59.6%). This may be due to the fact that the descriptions are quite long and being able to process them in both directions could lead to more relations discovered in the data, as opposed to using a single direction. Moreover, the lower performance of the Self-Attention model may be interpreted as the lack of enough data to train it properly.

As a further comparison, we explore the performance of the four models, measured with text-to-metaverse R@10, considering the length, i.e., the number of sentences, of the query description. The results are available in Table 2. As previously shown in Figure 3, most of the queries have between 10 and 19 sentences (10-19 range),

[11] https://github.com/aliabdari/Metaverse_Retrieval

**Table 1: Evaluation on the Test Set for the whole descriptions matching. *T-to-M* represents the text-to-metaverse task, whereas the opposite direction is captured by *M-to-T*. See Section 5.2 for the details.**

| | | Mean | Self-Attention | GRU | BiGRU |
|---|---|---|---|---|---|
| T-to-M | Recall@1 | 3.0 | 5.1 | 22.6 | **28.0** |
| | Recall@5 | 11.0 | 19.3 | 44.9 | **55.6** |
| | Recall@10 | 21.3 | 30.7 | 59.6 | **68.5** |
| | Median Rank | 32.0 | 23.0 | 7.0 | **4.0** |
| M-to-T | Recall@1 | 3.9 | 3.9 | 22.8 | **26.8** |
| | Recall@5 | 13.9 | 17.1 | 45.8 | **55.9** |
| | Recall@10 | 24.8 | 29.1 | 58.5 | **68.3** |
| | Median Rank | 28.5 | 24.0 | 7.0 | **4.0** |
| | NDCG@10 | 2.9 | 4.2 | 16.7 | **19.6** |
| | NDCG entire | 2.7 | 4.0 | 16.1 | **19.0** |

**Table 2: Performance of the four models considering the length of the query description. Details in Section 5.2.**

| | Text-to-Metaverse R@10 | | | | |
|---|---|---|---|---|---|
| Range | Mean | Self-Attention | GRU | BiGRU | # samples |
| 0-9 | 25.0 | 29.2 | 79.2 | **83.3** | *24* |
| 10-19 | 28.8 | 18.1 | 67.2 | **73.9** | *375* |
| 20-29 | 28.4 | 29.5 | 75.8 | **83.1** | *95* |
| 30-39 | 23.1 | 23.1 | **69.2** | 61.5 | *13* |
| 40-49 | **100.0** | 0.0 | 0.0 | 0.0 | *1* |

in which can be seen that the Bidirectional GRU model performs better than others; moreover, the Self-Attention performs worse than the simpler Mean pooling model. Considering the other ranges, it can be seen that there are situations in which the Self-Attention outperforms the Mean pooling (0-9, 20-29); the GRU performs better than its Bidirectional version in the 30-39 range, however, there are not enough samples (13) to make a conclusive statement.

## 5.3 Style, theme, and material queries

Looking for metaverses which follow a certain theme, e.g., a formal-looking scenario which is best suited for a business meeting, could be one of the more commonly used applications. To benchmark the performance of the considered models in a similar situation, we introduce a template for the queries which resembles what a user could have in mind. It has the following form: "I am looking for a scenario which follows a $x$ {style | theme | material}", where $x$ is an instance of either style, theme, or material, which represent respectively the style followed by the furniture (e.g., $x$ could be "Japanese" ones are often simple and minimalist, with earthy colors; whereas "Chinoiserie" use brighter colors, with shades of red and gold, and are decorated by more intricate patterns), the most prevalent type of patterns displayed on the furniture (e.g., "Lines" or "Striped Grid"), and the material used to build them (e.g., "Wood" or "Metal"). Overall, a total of more than 700 queries is obtained. In this experiment, since the query solely contains a single style, theme, or material, we consider a metaverse to be relevant to the query if at least half of its components satisfy the query. The results are reported in Table 3. The model based on the bidirectional GRU achieves the best results in retrieving the groundtruth, obtaining far better R@1 (19.3% compared to less than 8.4% of the GRU) and R@5 (70.8% compared to 60.3% obtained by the Self-Attention) than the other models. However, when considering the overall quality of the ranking lists, it can be seen that the nDCG reported by the Self-Attention is higher than the others (21.4% nDCG@10 and 24.2% nDCG compared to 15.0% and 20.3% of the bidirectional GRU). Notably, the queries used in this experiment are designed to be entirely different than the descriptions used at training time. Therefore, the results show that the model trained with descriptions performs reasonably well (e.g., R@1 of 19.3% in Table 3 compared to 28.0% in Table 1) even when using queries resembling human-written ones, showing a good generalization ability.

**Table 3: Evaluation of the style, theme, and material queries. More details in Section 5.3.**

| | Style/Theme/Material to Metaverse | | | |
|---|---|---|---|---|
| | Mean | Self-Attention | GRU | BiGRU |
| Recall@1 | 1.4 | 7.1 | 8.4 | **19.3** |
| Recall@5 | 50.1 | 60.3 | 48.5 | **70.8** |
| Recall@10 | 62.8 | 60.3 | 83.3 | **92.5** |
| NDCG@10 | 13.7 | **21.4** | 12.3 | 15.0 |
| NDCG entire | 20.4 | **24.2** | 19.8 | 20.3 |

## 5.4 Experiment about video queries

A promising application for a text-to-metaverse retrieval method is to facilitate the identification of metaverses featuring multimedia content associated with a particular individual, creating an immersive environment that evokes that person's life and work. For example, if we consider a painter like Van Gogh, we would expect to see reproductions of his paintings as well as elements that may have inspired him, such as flowers or fruits. In our setting, we used videos as the main multimedia element present in each metaverse. Therefore, in this experiment, the queries follow the form "I look for a room in which the TV showing $x$", where $x$ is the description of the content of a video. In this case, the relevant metaverses are those which contain the exact video mentioned in the query. Moreover, we altered the dataset by increasing the number of videos which are selected and distributed across the scenarios from 25 to 500. According to the results reported in Figure 6, it can be seen that with 25 videos the Mean and Self-Attention models are more effective at retrieving the metaverses which play a given video. As the number of videos increase, the performance of these two models decrease in favor of those based on GRU. It should be noted that the training procedure was not specifically optimized to improve the performance on this particular task, as its focus was on using descriptions, which contain multiple sentences about furniture and a single one about multimedia content, to retrieve metaverses. Therefore, it is possible that allocating more resources to this task (e.g., by giving more importance to the contents of the multimedia data in the description) could lead to better results.

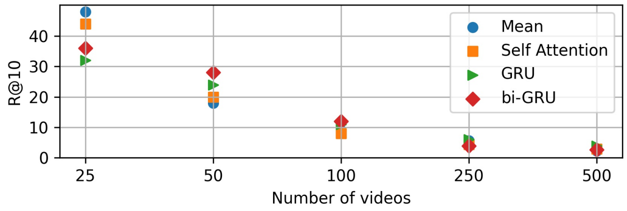

**Figure 6: Recall@10 on the test set using a variable amount of videos in the metaverses under analysis. More details in Section 5.4.**

## 6 CONCLUSION

With rapid technological progress in digital capabilities, an increasing interest is directed towards the metaverse and its many potential applications, ranging from entertainment and multimedia consumption, e.g., digital museums, to industrial and academic training, e.g., realistic looking environments where students and doctors can practice complex surgery procedures. Due to the ever-growing number of metaverses, there is an emergent need for advanced methods capable of retrieving the metaverses which best fit the users' needs. Therefore, in this paper, we introduced the novel task of text-to-metaverse retrieval, inspired by the recent advancements in cross-modal retrieval. To the best of our knowledge, such a research task has not been addressed either for metaverses or for 3D scenes, which can be seen as a less challenging scenario since they are not as multimedia-enriched as metaverse scenarios. In fact, the multimedia content present in the metaverse greatly influences its relevance to a user query, e.g., when trying to discriminate a modern art exhibition from an Impressionist one. To support research on this task, we collected a benchmark dataset consisting of more than 3000 metaverse-description pairs, since existing datasets do not contain multimedia-enriched scenes. Additionally, we developed a deep learning-based framework which we thoroughly tested with several experiments. In our experiments, we observed that using a bidirectional GRU on top of the sentence embeddings leads to better retrieval results when compared with three other models. When varying the type of query (style, theme, material query, or the query about the multimedia content) we observed a different behavior, which showed that also a simpler Mean pooling of the sentence embeddings can lead to a better quality of the ranking lists. Overall, both these results show that there is room for improvement, especially in the quality of the ranking lists, which needs to be significantly improved in order to achieve satisfactory results for the user.

Given the novelty of the task, there are several research directions which could be addressed. First, there is a lack of public datasets, which greatly limits the advancement in metaverse-related research. Therefore, collecting and releasing public datasets could be useful both for more general studies and for application-oriented ones, e.g., metaverses covering specific sectors such as art and entertainment. Second, in our setting, we relied on recent approaches based on deep Auto Encoders to extract the representation for the scenes. However, by representing the scene as a set of multiple views images, recent vision-and-language approaches [50] could be also used to solve this task: we reserve this research direction as future work. Moreover, our solution currently models the metaverse and its multimedia separately, possibly neglecting potential interactions between them. Therefore, future works should also investigate how to model them jointly.

## ACKNOWLEDGMENTS

This work was supported by the Department Strategic Plan (PSD) of the University of Udine–Interdepartmental Project on Artificial Intelligence (2020-25), MUR Progetti di Ricerca di Rilevante Interesse Nazionale (PRIN) 2022 (project code 2022YTE579), and by TechStar Srl, Italy.

## REFERENCES

[1] [n. d.]. SpaCy Universe. https://github.com/explosion/spaCy/blob/master/website/UNIVERSE.md. 17/11/2022.

[2] Hameed Abdul-Rashid, Juefei Yuan, Bo Li, Yijuan Lu, Song Bai, Xiang Bai, Ngoc-Minh Bui, Minh N Do, Heyu Zhou, Yang Zhou, et al. 2018. SHREC'18 track: 2D image-based 3D scene retrieval. *Training* 700 (2018), 70.

[3] Hameed Abdul-Rashid, Juefei Yuan, Bo Li, Yijuan Lu, Tobias Schreck, Ngoc-Minh Bui, Trong-Le Do, Mike Holenderski, Dmitri Jarnikov, Khiem T Le, et al. 2019. Shrec'19 track: Extended 2D scene image-based 3D scene retrieval. *Training (per class)* 700 (2019), 70.

[4] Giulio Paolo Agnusdei, Valerio Elia, and Maria Grazia Gnoni. 2021. A classification proposal of digital twin applications in the safety domain. *Computers & Industrial Engineering* 154 (2021), 107137.

[5] Sarah A Allman, Joanna Cordy, James P Hall, Victoria Kleanthous, and Elizabeth R Lander. 2022. Exploring the perception of additional information content in 360 3D VR video for teaching and learning. In *Virtual Worlds*, Vol. 1. Multidisciplinary Digital Publishing Institute, 1–17.

[6] Thomas Alsop. [n. d.]. Metaverse market size worldwide 2022-2030. https://www.statista.com/statistics/1295784/metaverse-market-size/.

[7] L. Ceci. 2022. Hours of video uploaded to YouTube every minute as of February 2020. https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute. [Online; accessed 31-March-2022].

[8] Yunpeng Chang, Zhigang Tu, Wei Xie, and Junsong Yuan. 2020. Clustering driven deep autoencoder for video anomaly detection. In *ECCV 2020*. Springer, 329–345.

[9] Long Chen, Wen Tang, Nigel W John, Tao Ruan Wan, and Jian J Zhang. 2020. Context-Aware Mixed Reality: A Learning-Based Framework for Semantic-Level Interaction. In *Computer Graphics Forum*, Vol. 39. Wiley Online Library, 484–496.

[10] Yuhao Cheng, Xiaoguang Zhu, Jiuchao Qian, Fei Wen, and Peilin Liu. 2022. Cross-Modal Graph Matching Network for Image-Text Retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.* 18, 4 (2022).

[11] Rajeswari Chengoden, Nancy Victor, Thien Huynh-The, Gokul Yenduri, Rutvij H Jhaveri, Mamoun Alazab, Sweta Bhattacharya, Pawan Hegde, Praveen Kumar Reddy Maddikunta, and Thippa Reddy Gadekallu. 2023. Metaverse for healthcare: a survey on potential applications, challenges and future directions. *IEEE Access* (2023).

[12] Hee-soo Choi and Sang-heon Kim. 2017. A content service deployment plan for metaverse museum exhibitions—Centering on the combination of beacons and HMDs. *International Journal of Information Management* 37, 1 (2017), 1519–1527.

[13] Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555* (2014).

[14] Gabriel Coelho, Luís Miguel Matos, Pedro José Pereira, André Ferreira, André Pilastri, and Paulo Cortez. 2022. Deep autoencoders for acoustic anomaly detection: experiments with working machine and in-vehicle audio. *Neural Computing and Applications* 34, 22 (2022), 19485–19499.

[15] Andrew Dawson et al. 2022. Data-driven consumer engagement, virtual immersive shopping experiences, and blockchain-based digital assets in the retail metaverse. *Journal of Self-Governance and Management Economics* 10, 2 (2022), 52–66.

[16] Nianchen Deng, Zhenyi He, Jiannan Ye, Budmonde Duinkharjav, Praneeth Chakravarthula, Xubo Yang, and Qi Sun. 2022. Fov-nerf: Foveated neural radiance fields for virtual reality. *IEEE Transactions on Visualization and Computer Graphics* 28, 11 (2022), 3854–3864.

[17] Jairo Díaz, Camilo Saldaña, and Camilo Avila. 2020. Virtual world as a resource for hybrid education. *International Journal of Emerging Technologies in Learning (iJET)* 15, 15 (2020), 94–109.

[18] Martin Engelcke, Dushyant Rao, Dominic Zeng Wang, Chi Hay Tong, and Ingmar Posner. 2017. Vote3deep: Fast object detection in 3d point clouds using efficient convolutional neural networks. In *ICRA*. IEEE, 1355–1361.

[19] Alex Falcon, Giuseppe Serra, and Oswald Lanz. 2022. A Feature-space Multimodal Data Augmentation Technique for Text-video Retrieval. In *30th ACMMM*. 4385–4394.

[20] Mohamed Habib Farhat, Xavier Chiementin, Fakher Chaari, Fabrice Bolaers, and Mohamed Haddar. 2021. Digital twin-driven machine learning: ball bearings fault severity classification. *Measurement Science and Technology* 32, 4 (2021), 044006.

[21] Huan Fu, Bowen Cai, Lin Gao, Ling-Xiao Zhang, Jiaming Wang, Cao Li, Qixun Zeng, Chengyue Sun, Rongfei Jia, Binqiang Zhao, et al. 2021. 3d-front: 3d furnished rooms with layouts and semantics. In *IEEE/CVF ICCV*. 10933–10942.

[22] Xiao Fu, Shangzhan Zhang, Tianrun Chen, Yichong Lu, Lanyun Zhu, Xiaowei Zhou, Andreas Geiger, and Yiyi Liao. 2022. Panoptic nerf: 3d-to-2d label transfer for panoptic urban scene segmentation. *arXiv preprint arXiv:2203.15224* (2022).

[23] Aidan Fuller, Zhong Fan, Charles Day, and Chris Barlow. 2020. Digital twin: Enabling technologies, challenges and open research. *IEEE access* 8 (2020), 108952–108971.

[24] Yusen Geng, Yuankai Zhang, Xincheng Tian, Xiaorui Shi, Xiujing Wang, and Yigang Cui. 2022. A novel welding path planning method based on point cloud for robotic welding of impeller blades. *The International Journal of Advanced Manufacturing Technology* 119, 11-12 (2022), 8025–8038.

[25] Jie Guan, Jay Irizawa, and Alexis Morris. 2022. Extended reality and internet of things for hyper-connected metaverse environments. In *IEEE VRW*. 163–168.

[26] Yun-Chih Guo, Tzu-Hsuan Weng, Robin Fischer, and Li-Chen Fu. 2022. 3D semantic segmentation based on spatial-aware convolution and shape completion for augmented reality applications. *CVIU* 224 (2022), 103550.

[27] Huy Ha and Shuran Song. 2022. Semantic abstraction: Open-world 3d scene understanding from 2d vision-language models. In *Conference on Robot Learning*.

[28] Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proc. of ICML*. 448–456.

[29] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM TOIS* 20, 4 (2002), 422–446.

[30] Geunho Jung and Sang Min Yoon. 2022. Monocular depth estimation with multi-view attention autoencoder. *Multimedia Tools and Applications* 81, 23 (2022), 33759–33770.

[31] A Sophia Koepke, Andreea-Maria Oncescu, Joao Henriques, Zeynep Akata, and Samuel Albanie. 2022. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia* (2022).

[32] Konrad Koniarski and Andrzej Myśliński. 2022. Feature Point Cloud Based Registration in Augmented Reality. In *Advances in Systems Engineering: Proceedings of the ICSEng 2021, December 14-16, Wrocław, Poland 28*. Springer, 418–427.

[33] Huilyung Koo. 2021. Training in lung cancer surgery through the metaverse, including extended reality, in the smart operating room of Seoul National University Bundang Hospital, Korea. *Journal of educational evaluation for health professions* 18 (2021).

[34] Heikki Laaki, Yoan Miche, and Kari Tammi. 2019. Prototyping a digital twin for real time remote control over mobile networks: Application of remote surgery. *Ieee Access* 7 (2019), 20325–20336.

[35] Hyun-Kyung Lee, Soobin Park, and Yeonji Lee. 2022. A proposal of virtual museum metaverse content for the MZ generation. *Digital creativity* 33, 2 (2022), 79–95.

[36] Lik-Hang Lee, Tristan Braud, Pengyuan Zhou, Lin Wang, Dianlei Xu, Zijun Lin, Abhishek Kumar, Carlos Bermejo, and Pan Hui. 2021. All one needs to know about metaverse: A complete survey on technological singularity, virtual ecosystem, and research agenda. *arXiv preprint arXiv:2110.05352* (2021).

[37] Xiang Li, Yuan Tian, Fuyao Zhang, Shuxue Quan, and Yi Xu. 2020. Object detection in the context of mobile augmented reality. In *IEEE ISMAR*. IEEE, 156–163.

[38] Zhixin Ling, Zhen Xing, Jiangtong Li, and Li Niu. 2022. Multi-level region matching for fine-grained sketch-based image retrieval. In *Proceedings of the 30th ACMMM*. 462–470.

[39] Kangcheng Liu, Zhi Gao, Feng Lin, and Ben M Chen. 2022. Fg-net: A fast and accurate framework for large-scale lidar point cloud understanding. *IEEE Transactions on Cybernetics* 53, 1 (2022), 553–564.

[40] Ying Liu, Lin Zhang, Yuan Yang, Longfei Zhou, Lei Ren, Fei Wang, Rong Liu, Zhibo Pang, and M Jamal Deen. 2019. A novel cloud-based framework for the elderly healthcare services using digital twin. *IEEE access* 7 (2019), 49088–49101.

[41] Siyu Lou, Xuenan Xu, Mengyue Wu, and Kai Yu. 2022. Audio-text retrieval in context. In *ICASSP 2022*. IEEE, 4793–4797.

[42] Avinash Madasu, Junier Oliva, and Gedas Bertasius. 2022. Learning to Retrieve Videos by Asking Questions. In *Proceedings of the 30th ACMMM*. 356–365.

[43] Nahuel A Mangiarua, Jorge S Ierache, and María J Abásolo. 2020. Scalable integration of image and face based augmented reality. In *Augmented Reality, Virtual Reality, and Computer Graphics: 7th International Conference, AVR 2020, Lecce, Italy, September 7–10, 2020, Proceedings, Part I 7*. Springer, 232–242.

[44] Bernard Marr. 2022. How Much Data Do We Create Every Day? The Mind-Blowing Stats Everyone Should Read. https://bernardmarr.com/how-much-data-do-we-create-every-day-the-mind-blowing-stats-everyone-should-read/. [Online; accessed 30-November-2022].

[45] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of IEEE/CVF CVPR*. 9879–9889.

[46] Thao Nguyen, Nakul Gopalan, Roma Patel, Matt Corsaro, Ellie Pavlick, and Stefanie Tellex. 2020. Robot object retrieval with contextual natural language queries. *arXiv preprint arXiv:2006.13253* (2020).

[47] Huansheng Ning, Hang Wang, Yujia Lin, Wenxi Wang, Sahraoui Dhelim, Fadi Farha, Jianguo Ding, and Mahmoud Daneshmand. 2021. A Survey on Metaverse: the State-of-the-art, Technologies, Applications, and Challenges. *arXiv preprint arXiv:2111.09673* (2021).

[48] Vaishakh Patil, Christos Sakaridis, Alexander Liniger, and Luc Van Gool. 2022. P3depth: Monocular depth estimation with a piecewise planarity prior. In *Proceedings of the IEEE/CVF CVPR*. 1610–1621.

[49] Pedro José Pereira, Gabriel Coelho, Alexandrine Ribeiro, Luís Miguel Matos, Eduardo C Nunes, André Ferreira, André Pilastri, and Paulo Cortez. 2021. Using deep autoencoders for in-vehicle audio anomaly detection. *Procedia Computer Science* 192 (2021), 298–307.

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

[51] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *IEEE CVPR*. 815–823.

[52] Luiz José Schirmer Silva, Djalma Lúcio Soares da Silva, Alberto Barbosa Raposo, Luiz Velho, and Hélio Côrtes Vieira Lopes. 2019. Tensorpose: Real-time pose estimation for interactive applications. *Computers & Graphics* 85 (2019), 1–14.

[53] Aziz Siyaev and Geun-Sik Jo. 2021. Towards aircraft maintenance metaverse using speech interactions with virtual objects in mixed reality. *Sensors* 21, 6 (2021), 2066.

[54] Wenshuang Song, Yanhe Gong, and Yongcai Wang. 2022. VTONShoes: Virtual Try-on of Shoes in Augmented Reality on a Mobile Device. In *IEEE ISMAR*. 234–242.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[56] Thang Vu, Kookhoi Kim, Tung M Luu, Thanh Nguyen, and Chang D Yoo. 2022. Softgroup for 3d instance segmentation on point clouds. In *Proceedings of the IEEE/CVF CVPR*. 2708–2717.

[57] Hao Wang, Xiang Bai, Mingkun Yang, Shenggao Zhu, Jing Wang, and Wenyu Liu. 2021. Scene text retrieval via joint text detection and similarity learning. In *Proceedings of the IEEE/CVF CVPR*. 4558–4567.

[58] Xuanyu Wang, Yang Wang, Yan Shi, Weizhan Zhang, and Qinghua Zheng. 2020. Avatarmeeting: An augmented reality remote interaction system with personalized avatars. In *Proceedings of the 28th ACMMM*. 4533–4535.

[59] Yabing Wang, Jianfeng Dong, Tianxiang Liang, Minsong Zhang, Rui Cai, and Xun Wang. 2022. Cross-Lingual Cross-Modal Retrieval with Noise-Robust Learning. In *Proceedings of the 30th ACMMM*. 422–433.

[60] Yilin Wang and Jiayi Ye. 2020. An overview of 3d object detection. *arXiv preprint arXiv:2010.15614* (2020).

[61] Lilong Wen, Yingrong Wang, Dongxiang Zhang, and Gang Chen. 2023. Visual Matching is Enough for Scene Text Retrieval. In *Sixteenth ACM WSDM*. 447–455.

[62] Michael Wray, Hazel Doughty, and Dima Damen. 2021. On semantic similarity in video retrieval. In *Proceedings of the IEEE/CVF CVPR*. 3650–3660.

[63] Bohan Wu, Suraj Nair, Roberto Martin-Martin, Li Fei-Fei, and Chelsea Finn. 2021. Greedy hierarchical variational autoencoders for large-scale video prediction. In *IEEE/CVF CVPR*. 2318–2328.

[64] Haitao Yang, Zaiwei Zhang, Siming Yan, Haibin Huang, Chongyang Ma, Yi Zheng, Chandrajit Bajaj, and Qixing Huang. 2021. Scene synthesis via uncertainty-driven attribute synchronization. In *Proceedings of the IEEE/CVF ICCV*. 5630–5640.

[65] Juefei Yuan, Hameed Abdul-Rashid, Bo Li, and Yijuan Lu. 2019. Sketch/image-based 3D scene retrieval: Benchmark, algorithm, evaluation. In *2019 IEEE MIPR*. IEEE, 264–269.

[66] Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards Automatic Learning of Procedures From Web Instructional Videos. In *AAAI Conference on Artificial Intelligence*. 7590–7598. https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17344