

Finding video shots for immersive journalism through text-to-video search

Lyndon Nixon
Modul Technology
Vienna, Austria
nixon@modultech.eu

Damianos Galanopoulos
CERTH-ITI
Thessaloniki, Greece
dgalanop@iti.gr

Vasileios Mezaris
CERTH-ITI
Thessaloniki, Greece
bmezaris@iti.gr

Abstract—Video assets from archives or online platforms can provide relevant content for embedding into immersive scenes or for generation of 3D objects or scenes. However, XR content creators lack tools to find relevant video segments for their chosen topic. In this paper, we explore the use case of journalists creating immersive experiences for news stories and their need to find related video material to create and populate a 3D scene. An innovative approach creates text and video embeddings and matches textual input queries to relevant video shots. This is provided via a Web dashboard for search and retrieval across video collections, with selected shots forming the input to content creation tools to generate and populate an immersive scene, meaning journalists do not need specialist knowledge to communicate stories via XR.

Keywords—video segmentation, video embedding, multimodal search, Generative AI, video to 3D models

I. INTRODUCTION

This paper presents an innovative approach to the discovery and integration of video assets into an immersive (XR) environment. Our use case is with journalists as part of the EU Horizon Europe funded TRANSMIXR project, however the same approach could be applied to any digital video collection.

The creation of an effective immersive experience involves the inclusion of digital objects and content assets into an immersive scene (a 3D space in which the user may move, whether overlaid over a camera view of the real world as in AR/MR or a fully computer-generated reality as in VR). Often XR content creators are generating digital assets directly for inclusion (such as 3D models of objects). However, especially if non-experts should be enabled to design and implement immersive content then it is vital that they have access to tools which permit them to identify relevant media, package related items and acquire content representations suitable for insertion into the scene (or for transformation into a representation), all in an environment which does not require XR expertise (unlike common content creation ecosystems such as Unity).

This is especially true in the case of journalists who can use XR to communicate a holistic view of a news story in an intuitive way. Usual tools and workflows for XR require additional training outside of the usual tasks in their job, a common barrier to usage in many occupations who could benefit from creating content in XR. In the case of the TRANSMIXR use case, we consider a journalist using XR to explore the various facets of a news story through the spatial placement of different digital objects and content assets. They need to be able to search available media relevant to a selected

news story, preview and select the assets they would re-use in the immersive scene and acquire an appropriate representation of those assets.

After considering the state of the art for immersive content creation for news in Section 2, we present the content creation workflow of the TRANSMIXR project in Section 3. We then turn to our innovative implementation of text-based search over video segments (Section 4) and their integration into an easy-to-use Web-based dashboard (Section 5) which make XR content creation (Section 6) accessible to non-experts such as journalists. We conclude with our outlook for more usage of XR by journalists (and other non-experts) in the future because of this work.

II. STATE OF THE ART

Immersive journalism may be defined as news “in a form in which people can gain first-person experiences of the events and situations described in news stories” [1]. There has been limited uptake by journalists and news studios of XR technology to communicate stories in an immersive manner. Issues have been the usability of XR content creation tools (by non-expert users such as journalists and news editors) and the lack of accessibility to suitable XR content assets (without specific creation from scratch). A limited number of examples can be found where XR was applied to communicate news stories, e.g. TIME launched an AR/VR app called TIME Immersive in July 2019 to highlight “visual journalism”¹, beginning with an experience of the Apollo 11 moon landing. However, the app is no longer available. Other recent examples include a virtual tour of the Notre Dame cathedral after the fire (BBC), crossing the Mediterranean with irregular migrants (The Guardian) or recreating the January 6 Capitol attack (Washington Post). However, due to the issues mentioned, such “immersive journalism” [2] appears still limited to a small number of specifically prepared experiences. Regardless, technologists remain convinced of the transformative potential of immersion in journalism, if these barriers to the creation of immersive journalistic experiences would be overcome [3].

XR content creation usually takes place within specialised software such as Unity. XR content assets can be created from scratch within these content creation environments and creations of other users may be shared (freely or licensed) via 3D asset sharing platforms. However, these materials may be too generic or inexact for news coverage, where accuracy is of vital importance. The emergence of Generative AI models has led to the capability of creating a 3D scene or 3D objects based on textual prompts, or from a set of images or video (e.g. an

¹ “TIME Launches New Augmented Reality and Virtual Reality App, TIME Immersive, to Showcase Groundbreaking Visual Journal

ism”. TIME. <https://time.com/5628880/time-immersive-app-ar-vr/> (accessed April 4, 2024).

object from all angles) [4]. While results are not yet perfect (and highly dependent on the quality of the input), this offers a potential way to create accurate representations (especially of real world objects) without specialist 3D modelling expertise. Work continues towards more photorealistic results [5]. Content can also be directly captured in an audiovisual representation. Both 360 degree and volumetric video capture produce audiovisual content that is directly applicable to an immersive environment, however the uptake of recording devices which capture these forms of video has been slow and therefore the availability of both 360 and volumetric video assets remains limited. As a result, the most ubiquitous form of audiovisual content available to content creators remains the ‘classical’ 2D linear video which is still a viable source of media for an immersive scene: either as input to a Generative AI model to produce a 360-degree scene or directly embedded onto a 2D plane in the 3D scene. If the entire scene has been captured in the audiovisual footage, even if as a 2D recording, a 360 or 3D (with depth) representation can even be constructed photorealistically with current tools, e.g. using Luma or Skybox AI. For generation of such a 360 or immersive scene, drone footage is of particular use [6] but generally needs to be created specifically on demand for the news story. Given the scale of availability of classical (2D linear) video material (consider alone online public video platforms such as YouTube) in comparison to specifically created video assets, we focus on the discovery of relevant media assets from existing video collections for the XR content creation as an aid to journalists who may not have the resources or budget to create new content for immersive journalism.

III. THE TRANSMIXR CONTENT WORKFLOW

The EU funded project TRANSMIXR is developing a data-driven XR content discovery, creation and delivery workflow (Fig. 1).

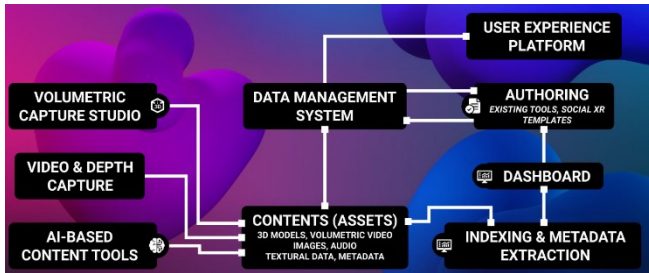


Fig. 1. High level view of the TRANSMIXR content workflow

As the lack of availability of suitable digital assets to form scenes and objects in an immersive experience is a major barrier to XR uptake, TRANSMIXR has focused on indexing publicly available Web content (textual news articles, social media, YouTube videos for example), content analysis to extract descriptive metadata, and a Web-based dashboard to browse the content by its metadata (an API is also available). The integration of video assets with their metadata extracted via content analysis into the dashboard has been done previously in an earlier project [7]. This content discovery part of the workflow can of course also be applied to any multimodal content sources, e.g. we index and analyse the newsfeed and video feed of our news partner AFP (available only to AFP content subscribers). The resulting content (digital assets) can either be used directly in the XR authoring

environment or transformed into an appropriate representation by Generative AI-based content tools.

To reiterate, the content discovery part of the TRANSMIXR workflow concerns three parts which can facilitate the subsequent immersive content creation:

1. Use of metadata analytics to identify topics of interest. For example, keyword analysis across news articles enables story detection, i.e. identify the news stories are in the focus of online reporting over time.
2. Direct retrieval of content to illustrate a topic. Both a Web-based dashboard and an API are available to query for content items via their metadata. Content items returned because of a search query are identified by a reference to their retrievable location (such as an URL if available on the Web).
3. Use of content items to generate new content for an immersive scene. Alternatively, the content which is found by the creator may be used as the input into a further content adaptation or generation step. For example, text or videos may be summarised to focus on their most salient parts for delivery in an immersive experience.

For the rest of the paper, we concentrate on the use of video material as content input to the immersive scene, which requires appropriate segmentation and annotation of the video content so that it can be associated with a selected news story.

IV. MAKING VIDEO FRAGMENTS FINDABLE

To make video assets available to search and retrieval systems, it is first necessary to analyse the video and output a machine processable description of the video's structure and semantics (descriptive metadata). A video analysis service developed in the TRANSMIXR project is accessed through a REST API and can be initiated by issuing a POST call, and the request JSON must include the video URL. The analysis of a submitted video is done in three main phases. First, the submitted video URL is downloaded and temporarily stored on the server. Then, the video is temporally segmented into scenes and shots, and finally, a feature extraction module analyses the video segments and extracts semantic features using cross-modal networks.

The video analysis service supports videos from the most popular video platforms, such as YouTube, Facebook and X, as well as videos hosted on file hosting platforms. The analysis of the videos starts on a temporal fragmentation level, and a video is first decomposed into scenes and shots. Scenes are temporally and semantically video segments that represent a story-telling part of a video. On the other hand, shots are segments which are a continuous capture by a single camera. A scene is typically a larger temporal fragment of a video and consists of one or more shots. We utilize the TransNet network [8] to temporally segment the video into scenes and shots. TransNet is a deep-learning architecture designed to detect shot boundaries in videos. Shot boundary detection involves identifying transitions between shots in a video sequence, such as cuts, fades, wipes, and dissolves. TransNet utilizes a convolutional neural network (CNN) architecture to analyze video frames and detect these transitions automatically. Finally, every video is decomposed into a sequence of video

shots, and from every shot, we extract keyframes with a sampling rate of 2fps.

After extracting shots' representative keyframes, the Text2Video module is utilized to extract the cross-modal visual semantic embeddings. The goal of the Text2Video module is to extract semantic information from both textual and visual data in the form of embeddings (multidimensional vectors) and to link text with video shots by representing those semantics in the same vector space and enables the free-text video search. For this, we follow a two-fold strategy. First, we utilize the TxV [9] network, a cross-modal network consisting of two key sub-networks designed for text-based video search. Secondly, inspired by the best-performing approach [21] and based on the conclusions of the TRECVID Ad-hoc video search 2023 [15], where the utilization of multiple cross-modal networks leads to improved performance, we also utilize numerous pre-trained cross-modal text-image networks as video-frame feature extractors. Apart from various pre-trained models of the vanilla CLIP [10] model, we also utilize pre-trained models from SLIP [11], BLIP [12], BLIP-2 [13], and LaCLIP [14].

TxV can capture complex semantic relationships between words and visual features, leading to more comprehensive and accurate representations of multimodal data. This network is designed to combine textual and visual features to develop multiple cross-modal common latent spaces. The TxV network is trained using a combination of four large-scale video caption datasets MSR-VTT [20], TGIF [18], ActivityNet [17] and Vatex [19]. Also, we trained multiple models using three learning rates (10^{-4} , $5 \cdot 10^{-5}$, 10^{-5}) and two optimizers (i.e., Adam and RMSprop), and we combined the results of these models to achieve the best possible performance. A tailored version of the TxV network is utilized in the Ad-hoc video search (AVS) 2023 benchmarking task [15][16] where it achieved Mean Extended Inferred Average Precision (MxinfAP) equal to 0.24.

After extracting the frame-level embeddings from the abovementioned text-image models, we develop shot-level embeddings by aggregating the frame embeddings using mean pooling. These generated embeddings are combined with the shot-level embeddings extracted by the TxV model. All of the generated vectors are returned as part of the service response. In order to learn how to combine all these embeddings, we follow a weighting aggregation approach at the query time, where each network is assigned with a unique weight value. During the training phase, we performed a grid search over these weights to determine their optimal values and evaluated our method's performance on the TRECVID AVS 2023 evaluation dataset, achieving MxinfAP equal to 0.299.

The analysed video shot collection can now be searched via textual query even though the extracted video shots do not have any associated textual metadata (at best, there is a title and description for the complete video only). When a user makes a textual query over the video shot collection, the Text2Video module encodes the query using the TxV network and all the aforementioned text-image models. Then, it calculates the similarity between these textual embeddings and the corresponding video shot embeddings using cosine similarity as a distance measure between vectors and develops similarity lists (one for each cross-modal network). Then, we utilize the learned optimal weighting values described above to aggregate the similarities between the user's query and all the shots from the video shot collection into a single list,

representing the final similarity. Fig. 2 below illustrates the components and workflow for the video analysis and retrieval.

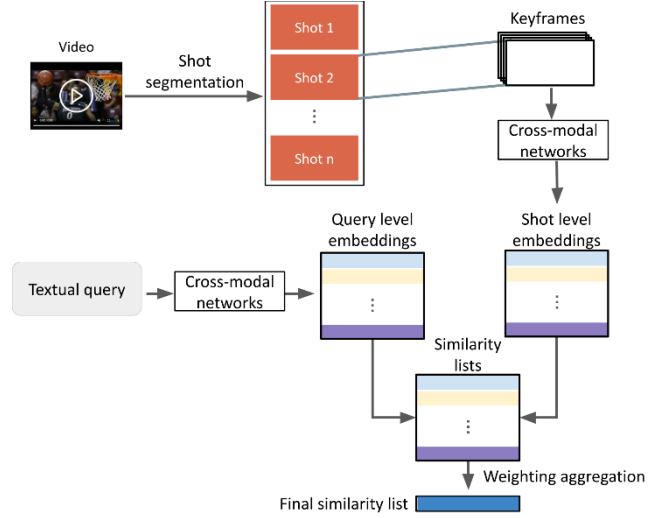


Fig. 2. Video analysis and retrieval system overview

V. DASHBOARD INTEGRATION

The textual search over video shots has been integrated into a Web dashboard which is online at <https://transmixr.weblyzard.com> (As AFP content cannot be made publicly accessible, a login is needed). Here, data sources can be selected which have been analysed and their descriptive metadata stored in an index such that content items may be found via textual search over that metadata. Unlike textual content, images and videos originally could only be searched based on the available textual metadata which was very restrictive and would not include details of individual objects and topics in the content (media titles and post descriptions as part of social media, for example). To make the video sources searchable in an effective manner by textual query, we call the Video Analysis service on every video ingested by our data collection pipeline (currently configured for public YouTube videos on predefined topics as well as AFP videos associated with news articles in their live newsfeed). The service response (video shots and their respective embedding vectors) is stored as metadata for each video item. The text2video search capability is integrated into the dashboard through a newly implemented service we named *embedify*. This service can take as input the text of a search query and determine the set of video shots which are semantically closest to the meaning of the text. To do this, we also convert the text query into a multidimensional vector embedding using the same approach as the Text2Video module. The video shots within the current search context of the dashboard (the selected data sources, the date and language settings) are sorted by cosine similarity between their embeddings and the text embedding of the search. The closest shots to the search query are returned.

An example can be given. A journalist explores recent news stories in the dashboard and identifies the topic of a solar eclipse in the global news coverage as an interesting subject for an immersive experience (after all, only people in the line of totality could truly experience the eclipse on April 8, 2024). They can check for relevant video material within

available data sources, having selected the story as their search query (in this case the terms “solar eclipse” and “totality” are used). Note how in this case they would not be able to “create” their own material for XR even if the budget and resources were available since the solar eclipse has already taken place and cannot be repeated (only in a fully virtual form, which would lack realism). Our sources are AFP video material (restricted to AFP members) and public YouTube video* (in a real news scenario, the uploader of the video should be contacted for permission for video re-use). In our case, we find 8 YouTube videos and 3 AFP videos. A part of the search results in the dashboard is shown in Fig. 3². By restricting the date range in the search to 8 April and afterwards, we can focus on video recordings of the actual eclipse, removing videos from before the event. We conclude, in this case, with 2 AFP and 2 YouTube videos that are available for immersive content creation.

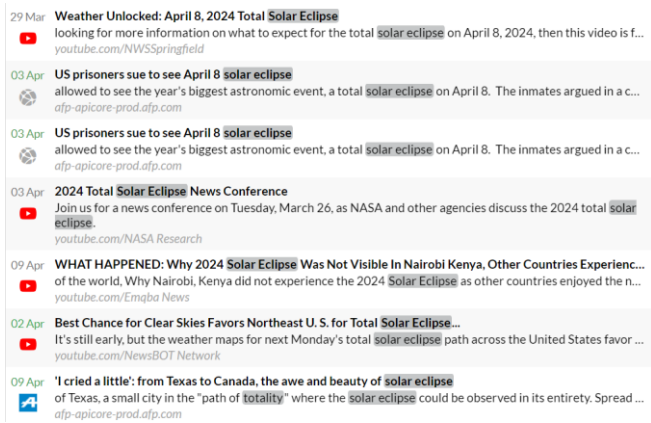


Fig. 3. Video results from AFP and YouTube sources for solar eclipse

Finding full videos related to the eclipse is not enough to move to the content creation step, developing the immersive experience. The journalist needs to drill down to the shots where there is clear coverage of the eclipse occurring, ideally not only showing the sky but also the surroundings (at ground level), rather than other content of these videos such as news anchors reporting the eclipse, interviews with eclipse spectators etc. Here we can make use of the Text2Video module to make the relevant video shots discoverable via textual search queries. Queries such as “sun”, “spectators” and “land” were used to find different shots (of the sun during the eclipse, of people watching the eclipse from the ground wearing protective eyewear, drone shots of the eclipse taking place over open land). Fig. 4 shows the advanced search with *embedify* in the dashboard, with a disjunctive query (Boolean OR) being used to select shots matching any of the textual queries within the selected set of (solar eclipse) videos.

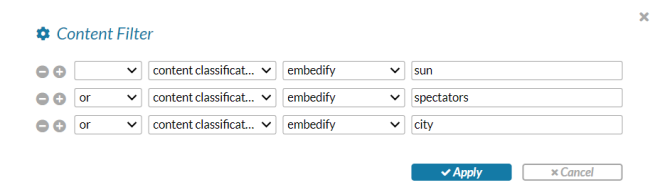


Fig. 4. Sample *embedify* search in the dashboard for relevant video shots

After a query is made, the matching video shots (those with a similarity to the search query within a certain threshold) are listed in the dashboard with a facility to preview them (through a play button, the video shot is played back within the Web interface) and “pin” those selected by the journalist for export (through the pin button, the URL of the video with a media fragment identifier for the specific shot is saved). Other content can also be exported from the dashboard: text from the video metadata (title and description) as well as images (keyframes extracted in the video analysis). These other modes of related content can, for example, be embedded into the immersive scene to provide more context around the chosen topic. Fig. 5 shows a sample video shot and image content discoverable within the dashboard.

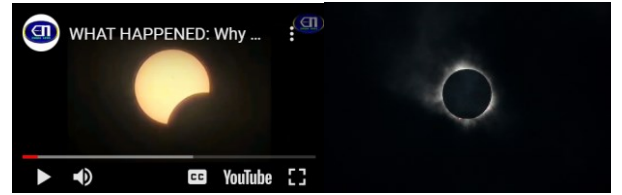


Fig. 5. Example content from the dashboard. (left) a video shot of the sun from a YouTube video, (right) a keyframe image from an AFP video.

At the end, the journalist has selected one suitable video shot of the solar eclipse in the sky and one suitable video shot of the solar eclipse taking place over a barren landscape in Antarctica. Having pinned both in the dashboard, their URLs are exported so that the video content is accessible to a XR content creation tool.

VI. XR CONTENT CREATION FROM VIDEO

We have experimented with Generative AI models for taking video shot inputs (or sets of frames from the shot) and generating 3D models or scenes of their content. While there are still limitations when the video material has not been specifically created for use in this form, both frames or continuous video can be effective when they (in totality) show the same object or scene from multiple angles at a good quality. Where the material proves insufficient for this, the images or video shots can of course still be placed on 2D planes in the immersive scene and are accessible in their original form.

In our case, the journalist could plan to create an immersive scene from the found video coverage, stitching the scene of the eclipse from above looking down with the scene of the eclipse from below looking up. As a proof of concept, we experimented with NVIDIA’s Neuralangelo [22], a Generative AI model which can construct a 3D scene out of classical 2D video footage. It claims to go beyond the state of the art in neural surface reconstruction, being better able to reconstruct the depth of scenes and objects filmed in 2D without needing a full multi-angle capture of the scene as is usual with current Generative AI models for video-to-3D (and which acts naturally as a major limiting factor in reuse of existing video assets in immersive content creation). Our video shot of the solar eclipse in the sky is a 15 second static

² The YouTube feed is a specially curated data collection from YouTube via its API, not a search over the whole of YouTube

shot (camera stays in a fixed position) recorded in timelapse mode. Our video shot of the solar eclipse over a barren landscape in Antarctica is also static, lasts 25 seconds and is also timelapse. In a situation where this content would be released to the public, we consider that an editor may first edit the videos to be the same duration and align their timelines (based on the time point where the eclipse starts to when it ends), so that the light conditions in both videos – the sky and the landscape – are synchronized. Neuralangelo produces a 3D mesh from the input videos which in turn may be rendered and converted into video format. The two 3D-rendered videos are stitched together and can now be explored as an immersive scene (the viewer can look around and observe the scene from different angles as well as move into or out of the space, while the solar eclipse is “taking place”). Fig. 6 provides a screenshot of this immersive scene.

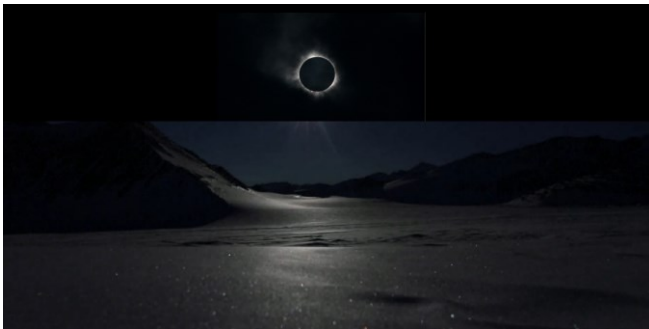


Fig. 6. Stitched video content of two 3D photorealistic scenes of the solar eclipse side by side

VII. CONCLUSIONS AND OUTLOOK

The use of XR to communicate current affairs and newsworthy events (as well as other topics) currently encounters technological and social barriers even though immersion can be very effective to enable consumers to explore topics from different angles or engage more deeply with the situation. In parallel to developments overcoming social barriers (e.g. cost of end user devices), we have presented a set of technologies and a tool that address technological barriers such as the difficulty to use XR content creation tools, lack of suitable XR content and missing integration of existing assets. Since there is a lack of appropriate resources for building photorealistic immersive experiences (legacy 3D object and scene creation has come from the computer gaming community, where results were intentionally unrealistic) and a cost and feasibility issue surrounding the specific creation of such resources directly (e.g. drone footage or recording a scene with a 360 video camera), we focus on the reuse of classical 2D video footage which is expected to be much more available and accessible to journalists (incl. the reuse of public user generated content with permission).

As part of an overall content workflow, we address the step of the discovery of available content which is a suitable input to a subsequent immersive content creation step. To make 2D video assets accessible, we analyze video files and apply video segmentation and content analysis. Innovatively, rather than an explicit annotation being the result of the content analysis (usually stored in a metadata document associated with the video file), we create a neural network-

based high dimensional embedding of each video shot. Through the Text2video module, we match those shot embeddings with a text embedding generated from a search query. This makes individual shots available to a textual search independent of descriptive metadata (which is typically missing from video segments). We showed how this Text2video is integrated into a Web-based dashboard so that journalists can search and browse shots intuitively, exporting the shots they find for input into an immersive content creation tool of choice.

Rather than dedicated XR environments such as Unity, we believe journalists and other non-experts will need much easier to use and intuitive tools for the creation of 3D objects as well as scenes. We believe Generative AI may prove to be the solution for getting from existing assets which can be found or created more easily (than assets which are directly applicable to XR) to 3D objects and scenes which make up an immersive experience of a news story or topic. We demonstrated an example of creating such an experience for the recent solar eclipse, finding and using two video shots from available data sources.

There is still a need for Generative AI which can produce good photorealistic 360 video, immersive scenes or 3D objects, especially from noisy input (images or videos which have not been created specifically for that purpose). This would revolutionize the ease of creating assets for XR. Given how rapidly this field is developing, there can be the expectation that we will draw closer to this with future model releases, with Neuralangelo a recent example of the emerging state of the art.

A combination of solutions for content discovery and for content creation may finally remove the current barriers to the use of XR for immersive journalism, and future consumers may be enabled to experience and explore the news in ways that we had not imagined, revolutionising our relationship with current affairs and global topics.

ACKNOWLEDGMENT

This paper has been funded by the EU Horizon Europe framework programme under the grant agreement 101070109 TRANSMIXR project (www.transmixr.eu).

REFERENCES

- [1] de la Peña, N., P. Weil, J. Llobera, B. Spanlang, D. Friedman, M. V. Sanchez-Vives, and M. Slater. 2010. “Immersive Journalism: Immersive Virtual Reality for the First-Person Experience of News.” *Presence: Teleoperators and Virtual Environments* 19 (4): 291–301.
- [2] Gynnild, A., Uskali, T., Jones, S., & Sirkkunen, E. (2020). What is immersive journalism?. *Immersive Journalism as Storytelling*, OAPEN open access library.
- [3] Wu, S. (2023). A Field Analysis of Immersive Technologies and Their Impact on Journalism: Technologist Perspectives on the Potential Transformation of the Journalistic Field. *Journalism Studies*, 24(3), 387-402.
- [4] Bai, S., & Li, J. (2024). Progress and Prospects in 3D Generative AI: A Technical Overview including 3D human. *arXiv preprint arXiv:2401.02620*.
- [5] Karnewar, A., Mitra, N. J., Vedaldi, A., & Novotny, D. (2023). Holofusion: Towards photo-realistic 3d generative modeling. In

- Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 22976-22985).
- [6] Pavlik, J. V. (2020). Drones, augmented reality and virtual reality journalism: Mapping their role in immersive news content. *Media and Communication*, 8(3), 137-146.
 - [7] Nixon, L., Apostolidis, E., Markatopoulou, F., Patras, I. and Mezaris, V. (2019): "Multimodal Video Annotation for Retrieval and Discovery of Newsworthy Video in a News Verification Scenario". *MultiMedia Modeling 2019 conference (MMM2019)*, Thessaloniki, Greece, January 2019.
 - [8] Jakub Lokoč, Gregor Kovalčík, Tomáš Souček, Jaroslav Moravec, and Přemysl Čech. 2019. A Framework for Effective Known-item Search in Video. In *Proceedings of the 27th ACM International Conference on Multimedia (MM '19)*. Association for Computing Machinery, New York, NY, USA, 1777–1785.
 - [9] Galanopoulos, D., Mezaris, V. (2023). Are All Combinations Equal? Combining Textual and Visual Features with Multiple Space Learning for Text-Based Video Retrieval. In: *Computer Vision – ECCV 2022 Workshops. ECCV 2022. Lecture Notes in Computer Science*, vol 13804. Springer, Cham.
 - [10] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., & Sutskever, I. 2021, Learning transferable visual models from natural language supervision. In *International conference on machine learning* (pp. 8748-8763). PMLR.
 - [11] Mu, N., Kirillov, A., Wagner, D., & Xie, S. (2022, October). Slip: Self-supervision meets language-image pre-training. In *European conference on computer vision* (pp. 529-544). Cham: Springer Nature Switzerland.
 - [12] Li, J., Li, D., Xiong, C., & Hoi, S. (2022, June). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning* (pp. 12888-12900). PMLR.
 - [13] Li, J., Li, D., Savarese, S., & Hoi, S. (2023, July). Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning* (pp. 19730-19742). PMLR.
 - [14] Fan, L., Krishnan, D., Isola, P., Katabi, D., & Tian, Y. (2024). Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36
 - [15] Awad, George, Keith Curtis, Asad A. Butt, Jonathan Fiscus, Afzal Godil, Yooyoung Lee, Andrew Delgado, et al. 2023. TRECVID 2023 - a Series of Evaluation Tracks in Video Understanding. In *Proceedings of TRECVID 2023*. NIST, USA.
 - [16] Galanopoulos, D., Mezaris, V. (2023), ITI-CERTH participation in AVS Task of TRECVID 2023, *Proceedings of TRECVID 2023*. NIST, USA.
 - [17] Caba Heilbron, F. et al. 2015. "ActivityNet: A Large-Scale Video Benchmark for Human Activity Understanding." In *Proc. Of IEEE CVPR 2015*, 961–70.
 - [18] Li, Y., Y. Song, L. Cao, J. Tetreault, et al. 2016. "TGIF: A New Dataset and Benchmark on Animated GIF Description." In *Proc. Of IEEE CVPR 2016*.
 - [19] Wang, X. et al. 2019. "Vatex: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research." In *Proc. Of IEEE/CVF ICCV 2019*, 4581–91.
 - [20] Xu, J., T. Mei, et al. 2016. "MSR-VTT: A Large Video Description Dataset for Bridging Video and Language." In *Proc. Of IEEE CVPR 2016*, 5288–96.
 - [21] J. He, R. Li, J. Guo, et al. (2023), Whu-Nercms At Trecvid2023 :Ad-Hoc Video Search (AVS) And Deep Video Understanding (DVU) Tasks, *Proceedings of TRECVID 2023*. NIST, USA.
 - [22] Li, Z., Müller, T., Evans, A., Taylor, R. H., Unberath, M., Liu, M. Y., & Lin, C. H. (2023). Neuralangelo: High-fidelity neural surface reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8456-8465).