SHIFT

MetamorphoSis of cultural Heritage
Into augmented hypermedia assets
For enhanced accessibiliTy
and inclusion

Funded by
the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Document info

Document ID:	D2.2 - Automatic generation of motion sequences from pictorial reposiroties – final version
Version date:	31.03.2025
Total number of pages:	34
Abstract:	This deliverable reports on the research activities that have been carried out for the transformation of cultural content to become appealing and interactive using computer vision algorithms. The research activities in the WP reflect upon the latest development in the field of artificial intelligence powered by deep-learning models to enable digital content transformation. In this deliverable three main technologies have been reported including the segmentation of background and foreground objects. The use of stable diffusion models in the context of generative video models that has been applied across a range of case-studies. Despite several examples being included in literature, the study of video generative models for revitalising historical artwork within the context of cultural heritage assets has not been sufficiently addressed within these studies. The research study aims to demonstrate the effectiveness of two generative video models namely (i) the video synthesis model and (ii) the stable video diffusion model in creating content that would be acceptable for cultural heritage institutions. Lastly, the deliverable also reports on the activities related to the extraction of image and video descriptions from the transformed paintings.
Keywords	Computer vision toolkit, content enrichement, stable diffusion, neural networks, diffusion models, 3D reconstruction, super resolution, action sequence library

Authors

Name	**Organisation**	**Role**
Krishna Chandramouli	QMUL	Researcher
Ebroul Izquierdo	QMUL	Researcher
Tomas Piatrik	QMUL	Researcher
Sokratis Nifakos	MDS	Researcher
Andrea Belin	MDS	Researcher

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Allister Carter	MDS	Researcher

REVIEWERS

Name	Organisation	Role
Ioana Crihana	ANBPR	Reviewer
Razvan Purcarea	SIMAVI	Reviewer

VERSION HISTORY

Version	Description	Date
0.1	Table of contents distributed among partners	09/12/2024
0.2	First round of inputs from partners	23/01/2025
0.3	Second round of inputs	14/02/2025
0.4	Final consolidated and release of deliverable draft	31/03/2025
0.5	Final revisions and comments addressed	31/03/2025

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

EXECUTIVE SUMMARY

The deliverable titled "Automatic generation of motion sequences from pictorial repositories (D2.2)" reports on the activities that were carried out since the publication of the previous deliverable (D2.1) in M12. The overarching goal of the WP is to facilitate the cultural heritage institutions to improve the appeal of the cultural content and achieve higher degree of user engagement through the adoption of digital transformation strategies. To this end, the activities of this work package are logically constructed that progressively aims to improve the content appeal and enrich the accessibility of the cultural assets using digital technologies.

Building up on the previous research activity that was carried out to study the user perception of cultural assets such as paintings, the information is then used to model the digital transformation of paintings into meaningful visual sequences. To achieve this objective, in this deliverable we report on the use of generative video models that are based on stable diffusion models. The effectiveness of the generative video models to digital transformation has been tested and evaluated against a series of paintings with different cultural contexts.

Additionally, the deliverable also reports on the extensive experimentation that were conducted towards the image foreground and background subtraction and to generate a set of action verbs from the analysis of the paintings.

The research outcomes reported in the deliverable will be used to produce digitally transformed content to increase the appeal of the cultural heritage assets. Following a summary of results in published in this deliverable a systematic methodology for the overall performance of the generated content will be published within WP5.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Content

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

LIST OF FIGURES

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Abbreviations and Acronyms

Abbreviation / Acronym	Description
AI	Artificial Intelligence
AOI	Area of interest
BC	Boundary condition
CCI	Creative and cultural industry
CHI	Cultural Heritage Institutions
CMVS	Clustering Multi-View Stereo
CNN	Convolutional neural network
CogVLM	Cognitive Visual Language Model
CPINN	Conservative PINN
CRF	Conditional Random Fields
DGM	Deep Galerkin Method
DRM	Deep Ritz method
FCN	Fully connected network
GANs	Generative Adversarial Network
MVS	Multi-View Stereo
PBR	Physical based rendering
PCNNs	Physics-constrained neural network
PDEs	Partial Differential Equations
PINNs	Physics-informed Neural Network
RCNN	Region based (-with) convolutional neural network
SAM	Segment Anything Model
SfM	Structured from Motion
SR	Super Resolution
SVD	Stable Video Diffusion
SVM	Support Vector Machines
TLS	Terrestrial Laser Scanning

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

ViT	Vision Transformers
VR	Virtual Reality
WP	Work package
YOLO	You Only Look Once

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

1. INTRODUCTION

One of the recent developments in the field of Artificial Intelligence (AI) is an increase in popularity of AI-generated content (AIC), which is considered one of the most prominent research fields [1]. The interest is not only a result of extensive attention and scholarly investigation, but it has also exerted profound influence across industries and other applications such as computer graphics, art, and design. Extending from the development of image synthesis [2], research focus around AIC has shifted towards video generation and synthesis [3]- [5]. One of the driving factors for the development of new models in video synthesis and generation is to facilitate the creation of video as part of training activities resulting in the simulation of scenarios which supports and enables real-world decision making [6]. Similarly, many such examples are being identified to support the need for developing video generation.

Despite the progress being made in the field of video generation, one area of study and research that has not gained much attention is cultural heritage. One of the biggest challenges in dealing with cultural heritage is to ensure cultural integrity is maintained, and that generated video sequences do not diminish the quality and aesthetics of the cultural imprint left behind by the artist. While history and cultural heritage need to be preserved in their original form, the birth of "digital natives" has led to cultural heritage institutions investing and developing mechanisms to encourage user engagement. Since the dawn of the content economy, user engagement has been placed at the heart of digital transformation strategies being adopted within the creative and cultural industry's (CCI) engagement with citizens. The trend is complemented by the changing demographics within Europe, where increasing number of citizens and users of content economy demand high-quality content. An increasing number of cultural heritage institutions are beginning to adopt digital transformation strategies, and the rate of adoption has been expedited since the COVID-19 pandemic. To enable cultural heritage institutions to effectively attract and engage with citizens, it is vital to "revitalise" the cultural heritage content that will serve communities. To achieve this goal, in this paper a study of video generative models for cultural heritage content is presented, that would focus on maintaining the historical context while enriching the user experience in interacting with the content.

1.1. Scope and objectives

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

The aim of the work package is to implement advanced AI tools and algorithms that enable modality transformation of the content from pictorial representation of cultural artefacts (such as paintings) into meaningful short video clips. The short clips that need to be generated will be associated to the cultural and historically accurate forms of representation, while at the same time enriching the content to engage the audience of the cultural heritage institutions. To accomplish this complex objective, the workpackage (WP) will implement a sequence of algorithms that will progressively enable the processing of pictorial content to be transformed into short motion sequences.

To achieve this objective, in consultation with the end-users, four forms of cultural assets have been studied and these include (i) portrait paintings; (ii) landscape representations of the city or a fictitious representation of the natural regions; (iii) historical and religious scenes that have been depicted by several artists over the years; and (iv) medical images depicting the representation of medical practices.

Additionally, the work package activities have also considered the data sets that have been provided by the end-users of SHIFT to be further processed, which are described in detail below.

1.2 Recent advances in the computer vision toolkits

The aim of the paper is to conduct a study on the effectiveness of the video generative models for creating content while maintaining the cultural integrity. To achieve this purpose, two models have been selected namely (i) stable video diffusion (SVD) (Blattmann et al., 2023) and (ii) I2VGen-XL, an image to video synthesis model using cascaded diffusion models (Zhang et al., 2023). As the intention of the study is to review the performance of the models when applied as is, in the application of both models, the pretrained models have been selected. In the case of the SVD no additional input parameters are required apart from the painting. In the case of the I2VGen model, a list of prompts has been used to generate the video synthesis results. As an evaluation of the project results, the aesthetics of the generated video sequences depicting the video texture is reviewed. As there is no baseline evaluations available for the evaluation of the quality of the cultural paintings being transformed to video sequences, only the use of qualitative metrics have been considered, which is presented in Section IV.

1.3 SHIFT computer vision toolkit

The topic of transforming the static images into video sequences has attracted several previous researchers to conduct initial assessments and research in this

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

area. The field of study in which the transition was presented and carried out was referred to as "producing looping videos", known variously as "video textures, cinemagraphs, or live photos". These techniques typically take as input a longer video sequence and, through some analysis of the motion, produce a single seamlessly looping video, or an infinite (yet not obviously looping) video (Schödl et al., 2000). The authors state that, the term cinemagraph often refers to selective animation of looping clips, where only certain parts of the frame, as chosen by the user, are animated (or de-animated). Newer approaches such as those reported in (Bhat et al., 2004; Chuang et al., 2005; Holynski et al., 2020; Lin et al., 2019) perform this task fully automatically, determining which regions are easily looped, and which regions contain motions that are large in magnitude or otherwise unsuitable for looping. These approaches have also been extended to operate on specific domains, such as videos of faces (Bai et al., 2012), urban environments (Yan et al., 2016), panoramas (Agarwala et al., 2005), and continuous particle effects (Agarwala et al., 2005; Lin et al., 2019). It is noted that all the methods proposed require a video as input to process the motion sequences. Complementary to the previous research topics, there are also a number of methods aimed at animating still images. Recently, these techniques have gained popularity through commercial applications such as Plotagraph and Pixaloop , which allow users to manually "paint" motion onto an image. Further complementary to the previous two approaches on generating motion sequences, Holynsk et al 2020 (Holynski et al., 2020) focusses on approaches to perform the selection of certain spatial regions to be selected automatically.

- Physical simulation. Instead of manually annotating the direction and magnitude of motion, the motion of certain objects (such as boats rocking on the water) can be physically simulated, as long as each object identity is known and its extent is precisely defined (Chuang et al., 2005) or automatically identified through class-specific heuristics (Jhou & Cheng, 2016). Since each object category is modelled independently, these methods do not easily extend to more general scene animation.
- Using videos as guidance. Alternatively, motion or appearance information can be transferred from a user provided reference video, containing either similar scene composition (Prashnani et al., 2017), aligned information from a different domain, such as semantic labels (Wang et al., 2019), or unaligned samples from the same domain (Cheng et al., 2020; Siarohin et al., 2020). Instead of a single user provided video, a database of homogeneous videos can be used to inherit nearest-neighbour textural motion, assuming a segmentation of the dynamic region is provided (Okabe et al., 2009).

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

1.4 Content specification for processing

For the study, it was important to consider what forms of video sequence generation from the paintings would be relevant by the cultural heritage institutions. Therefore, the following constraints were considered in evaluating the video sequences:

- The effects created in the video sequence should be natural
- The colour palette of the video sequence should align to the original painting.
- The historical context of the painting should not be altered
- No new information should be introduced in the painting that would denigrate the original artistic impact of the painting
- No artificial objects or items that are not a part of painting should be introduced.

For conducting the experiments, a total of 11 historical paintings from the 18th to 19th century were considered as depicted in Figure 5. The selection of the paintings was carried out based on the following constraints:

- A historical painting that depicts landscapes in general
- The depiction of natural elements in the paintings such as water, cloud, plains, trees, mountains among others.
- The possibility of content enrichment with motion embedded within the painting.
- Wide range of colours and colour palette depicting the painting
- A varying degree of themes within landscape

1.5. Structure of the report

The rest of the deliverable is structured as follows. In Section, a review of the background and foreground subtraction implementation for the generation of digital representation of cultural assets is presented. This section reports on the different algorithms that have been considered for the experimentation. In Section 3, the research activity related to generative video models is presented. In particular, two state of the art AI models have been selected to be applied in the context of transforming digital content from paintings into video sequences. The experimental results presented in the section showcase the effectiveness of one model against the other. In Section 4, a brief report on the creation of the action vocabulary is presented, followed by conclusion in Section 5.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

2. BACKGROUND AND FOREGROUND SUBTRACTIONS

Cultural heritage preservation has traditionally relied on physical conservation and restoration techniques. However, with the advancements in artificial intelligence (AI) and computer vision, digital reconstruction methods have gained prominence. By leveraging machine learning models, 3D imaging, and virtual reality (VR) platforms, cultural artifacts can now be digitized, reconstructed, and displayed in interactive virtual spaces.

The introduction of AI-powered solutions for cultural heritage preservation offers an innovative way to digitize, restore, and archive historical artifacts. Traditional methods often rely on manual restoration, photography, and documentation, which can be time-consuming and prone to human error. AI-enhanced methodologies, on the other hand, introduce automated segmentation, enhanced depth perception, and sophisticated 3D rendering techniques that drastically improve efficiency and accuracy.

Some of the primary AI-driven techniques used in the WP2, D2.1/D2.2 include:

- Segmentation using Deep Learning: Allows the accurate extraction of artifacts from complex backgrounds.
- Monocular Depth Estimation: Predicts depth from single 2D images, essential for generating 3D structures.
- Structure-from-Motion (SfM): Constructs a 3D point cloud from multiple 2D images.
- Multi-View Stereo (MVS): Enhances the density and accuracy of the 3D reconstruction.
- Poisson Surface Reconstruction: Converts point clouds into detailed 3D meshes.
- VR/AR Rendering: Provides interactive, immersive access to cultural heritage assets.

The reconstruction of historical artifacts from 2D images requires a combination of AI-based segmentation, depth estimation models, photogrammetry techniques, and 3D rendering algorithms. Each step in this pipeline is critical for preserving fine details and ensuring realistic virtual interactions. Below, we describe each phase in detail.

The AI models tested in this case were essential in facilitating the transformation of 2D images into detailed 3D models in order to perform digital reconstructions of historical artifacts and environments. These models were integrated into FORTH's VR Museum, where segmented objects from historical datasets were used for interactive VR exhibitions. The goal of this transformation was to preserve and

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

enhance the accessibility of cultural heritage by digitizing artifacts in a format that allows users to engage with them in a realistic and immersive manner.

Transformation Process: From 2D to 3D

The transformation process involved multiple steps, each employing specialized deep-learning models and computer vision techniques to ensure the fidelity and accuracy of the reconstructed artifacts. These models and techniques will be discussed in details in other sections.

Segmentation and Foreground Extraction

One of the initial and most critical steps in the transformation process was the segmentation of historical artifacts from their backgrounds. To achieve this, we utilized SAM (Segment Anything Model), a deep-learning-based segmentation model trained on very large datasets of images. This model was designed to generalize across different types of images without requiring task-specific training. The segmentation process involved several stages. First, the raw image was pre-processed by applying histogram equalization and noise reduction techniques to improve the clarity of the objects. The SAM model was then applied to extract the foreground artifact by detecting object boundaries and generating an accurate segmentation mask. To further refine the segmentation, post-processing steps such as morphological filtering were applied to eliminate noise and small unwanted artifacts. This resulted in a clean, isolated representation of the object that could be used for the subsequent reconstruction steps.

The benefit of this model is that it is designed to generalize across different types of images without requiring task-specific training, thus making it more applicable in a wider range of scenarios.

SAM consists of an image encoder and a promptable mask decoder. We can break down the process as follows:

Image Encoding: Given an input image, SAM uses a Vision Transformer (ViT) backbone to compute an embedding:

Depth Estimation and 3D Structure Reconstruction

With the segmented artifacts isolated, the next challenge was to infer their depth information, and this is a very important step in converting a 2D image into a 3D representation. This was accomplished using a monocular depth estimation model based on an encoder-decoder convolutional neural network (CNN). The model was pre-trained on large-scale depth datasets, enabling it to predict depth maps for single input images with high precision.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

The predicted depth maps were post-processed using bilateral filtering to smooth out noise while preserving edges. Moreover, a global optimization algorithm was applied to test depth consistency across multiple viewpoints. This step gave us the opportunity for a more realistic approximation of the three-dimensional structure of the artifacts and functioned as the baseline for the next phase of reconstruction.

Point Cloud Generation Using Structure-from-Motion (SfM) and Multi-View Stereo (MVS)

Point Cloud Generation using Structure from Motion (SfM) and Multi-View Stereo (MVS) is a sophisticated approach within the field of computer vision aimed at reconstructing three-dimensional (3D) models from two-dimensional (2D) image datasets. SfM estimates camera motion and generates sparse point clouds by analyzing multiple overlapping images, while MVS enhances these outputs to create dense, detailed 3D representations. This integrated process is crucial for applications in robotics, augmented reality, cultural heritage documentation, and urban modeling, where accurate spatial understanding is essential. The significance of SfM and MVS lies in their ability to transform vast collections of images into actionable 3D data, facilitating advancements in various domains, including archaeological research, environmental monitoring, and architectural visualization. The methods enable researchers and professionals to visualize and analyze complex structures and scenes with high fidelity, thus contributing to fields such as cultural heritage preservation and urban planning. Despite their transformative potential, the integration of SfM and MVS is not without challenges. Issues such as computational intensity, the quality of initial image datasets, and the handling of occlusions and lighting variations can impact the accuracy and efficiency of 3D reconstructions. Furthermore, recent developments in machine learning are prompting discussions about their implications for the future of point cloud generation, especially regarding automation and real-time processing capabilities. In summary, the combination of Structure from Motion and Multi-View Stereo represents a pivotal evolution in 3D reconstruction methodologies, driving innovation across various sectors while also presenting technical challenges that researchers continue to address. As the technology advances, it holds the potential to further revolutionize how we capture and interact with the three-dimensional world around us.

Structure from Motion (SfM)

Structure from Motion (SfM) is a photogrammetric range imaging technique that aims to estimate three-dimensional structures from a sequence of two-dimensional images, often coupled with local motion signals. This method serves as a significant challenge in the fields of computer vision and visual perception, focusing on reconstructing static scenes based on the estimation of camera motion corresponding to these images

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Multi-View Stereo (MVS)

Multi-View Stereo (MVS) is a critical technique in computer vision that reconstructs dense 3D geometries of objects or scenes using multiple calibrated 2D images captured from various angles. MVS plays a pivotal role in applications such as robotics, augmented and virtual reality, and automated navigation, allowing for a comprehensive understanding of spatial environments

Integration of SfM and MVS

Structure from Motion (SfM) and Multi-View Stereo (MVS) are integral components in the 3D reconstruction pipeline, each serving distinct but complementary roles. SfM primarily focuses on estimating camera parameters and creating a sparse 3D model from a series of 2D images, while MVS refines this output by generating a dense 3D reconstruction.

Process Overview

The process begins with the SfM algorithm, which utilizes multiple images to output both the camera parameters for each image and a coarse representation of the scene, known as the sparse reconstruction. This output is crucial for the subsequent MVS phase, which takes advantage of the known camera parameters to resolve ambiguities in 3D vertex computation. MVS operates by examining image regions along epipolar lines, leveraging the geometry established by SfM to match points more efficiently than traditional feature-based methods, particularly in areas where descriptor matching is challenging due to lighting or material variations.

Enhancements and Variants

To enhance the performance of MVS, especially in handling larger datasets, additional algorithms like PMVS2 (Patch-based Multi-View Stereo) and CMVS (Clustering Multi-View Stereo) have been developed. PMVS2 focuses on refining the mesh produced by SfM, producing a denser output that fills in gaps left by the initial sparse model. However, it can be computationally intensive and memory-demanding, particularly for extensive image sequences. CMVS addresses these limitations by clustering the initial sparse 3D outputs, allowing PMVS2 to process each cluster separately, thus improving execution speed and memory efficiency. The integration of artificial intelligence and machine learning into SfM and MVS holds promise for the future, aiming to improve accuracy and efficiency in real-time applications. Machine learning techniques could enhance the interpretation of complex scenes, enabling better object differentiation and real-time 3D model refinement. This integration is poised to transform fields such as augmented reality and autonomous navigation, pushing the boundaries of what SfM and MVS can achieve in practical applications

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Data Acquisition

Data acquisition in the context of Point Cloud Generation using Structure from Motion (SfM) and Multi-View Stereo (MVS) involves the systematic gathering of images and associated data to create detailed 3D models. This process can leverage various techniques and technologies, including pre-trained imagery models, aerial imagery, and terrestrial laser scanning (TLS).

Mesh Reconstruction and Texturing

This section refers to the technological processes and methodologies employed to digitally capture, recreate, and represent historical artifacts and sites. This field has gained prominence as a vital aspect of cultural heritage preservation, enabling more precise documentation and broader accessibility to heritage assets through advanced techniques such as three-dimensional (3D) scanning, photogrammetry, and mesh processing. The growing integration of these technologies into cultural heritage practices reflects a significant evolution in how heritage is understood, shared, and safeguarded for future generations. The importance of mesh reconstruction and texturing lies in its ability to produce detailed, photorealistic digital representations of artifacts, which facilitate scholarly research, public engagement, and educational outreach. However, this evolving field is not without its challenges. Ethical concerns regarding the provenance of artifacts and the impact of digital surrogates on the engagement with original materials raise questions about the implications of digital representation. Moreover, technical difficulties related to accuracy, data interpretation, and funding limitations can hinder the effectiveness of these processes, emphasizing the need for continuous research and standardization within the sector. As cultural heritage institutions grapple with these issues, the future of mesh reconstruction and texturing remains intertwined with technological advancements and collaborative efforts across disciplines. Ultimately, mesh reconstruction and texturing serve as critical tools in the preservation of cultural heritage, enabling a richer understanding and appreciation of the past while addressing contemporary challenges in accessibility and representation. As this field continues to evolve, its influence on cultural heritage practices will likely expand, promoting sustainability and inclusivity in the stewardship of historical assets.

In the SHIFT project, once a detailed point cloud was generated, the next essential step was to convert it into a continuous 3D surface. A raw point cloud consists of discrete data points that outline the structure of an object, but to create a fully immersive VR experience, these points needed to be transformed into a high-resolution, continuous surface. This was accomplished through Poisson Surface Reconstruction, a powerful technique that interpolates the structure of the object by estimating a smooth, dense mesh from the scattered point data.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Poisson Surface Reconstruction works by assuming that the sampled points approximate a surface and that their normals represent a function's gradient field. Using these assumptions, it computes a watertight, dense mesh that captures the overall structure while maintaining high geometric fidelity. This method proved particularly effective for cultural heritage artifacts, as it preserved fine surface details while avoiding excessive noise and irregularities often present in raw point cloud data.

After generating the base mesh, it underwent further refinement using Delaunay Triangulation, an optimization technique designed to ensure uniformity in the triangular facets that compose the mesh. By restructuring the topology and eliminating overlapping or stretched triangles, this method enhanced structural integrity, making the models both visually accurate and computationally efficient for rendering in VR environments.

To achieve photorealistic rendering, high-resolution textures were applied to the reconstructed mesh. The texture mapping process involved generating UV coordinates, which allowed the original 2D image textures to be accurately wrapped around the 3D model. This step was crucial in preserving the artifact's realism, as it ensured that surface details, such as engravings, color variations, and material characteristics, were faithfully transferred to the digital model.

The final phase of texturing involved texture blending and optimization. Multiple texture images captured from different viewpoints were stitched together to create a seamless, high-resolution surface representation. This process included color correction, normal mapping, and shadow preservation, ensuring that the artifacts appeared natural when displayed in varying lighting conditions within the VR environment. By combining Poisson Surface Reconstruction, Delaunay Triangulation, and UV mapping, the SHIFT project successfully produced high-quality, interactive 3D models that could be explored in a virtual museum setting. The enhanced realism of these models improved user immersion, making it possible to study, manipulate, and engage with historical artifacts in ways that were previously impossible through traditional museum displays.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Figure 1 - Physical objects

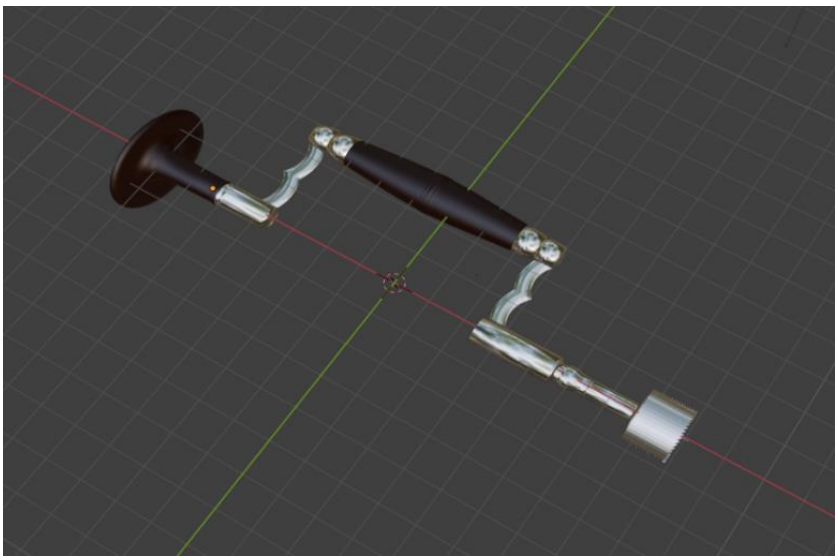

Figure 2 - Digital representation of the physical objects

As part of WP2 work, multiple historical artifacts were extracted, processed, and reconstructed as 3D models for use in VR environments. The image in this part is an example the extracted objects, showcasing the effectiveness of the depth estimation, segmentation, and reconstruction techniques applied in the project. he extracted 3D model exhibits a high degree of accuracy in both structural representation and material detailing. The model successfully retains fine geometric details, particularly in complex, curved, and articulated sections of the artifact. The smooth surface transitions were made possible through Poisson Surface Reconstruction and Delaunay Triangulation, while high-resolution texture

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

mapping ensures the visual authenticity of the final VR representation. The structural accuracy of the models demonstrates a consistent and well-defined topology, ensuring that the VR representation closely matches the real-world artifact. The segmentation process effectively isolated the artifact from the background, allowing for clean and precise mesh reconstruction. The texture quality was also carefully preserved, as the UV mapping process enabled accurate wrapping of textures, minimizing distortions in detailed regions. Color fidelity and material properties were well maintained, ensuring that the object appeared realistic within a VR environment. Optimization for VR integration was another important aspect, as the mesh density was adjusted to balance detail preservation and rendering performance. The model was exported in standard VR-compatible formats, such as glTF and FBX, to ensure smooth integration into platforms like Unity and Unreal Engine. Some areas for improvement include further refinement of edges to minimize aliasing artifacts that may still be present. Enhanced physically based rendering (PBR) shading techniques could improve realism under different lighting conditions in VR. Additional texture blending techniques could be applied to further improve uniformity across large surfaces, reducing visible seams between texture maps.

Figure 3 - Physical object

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

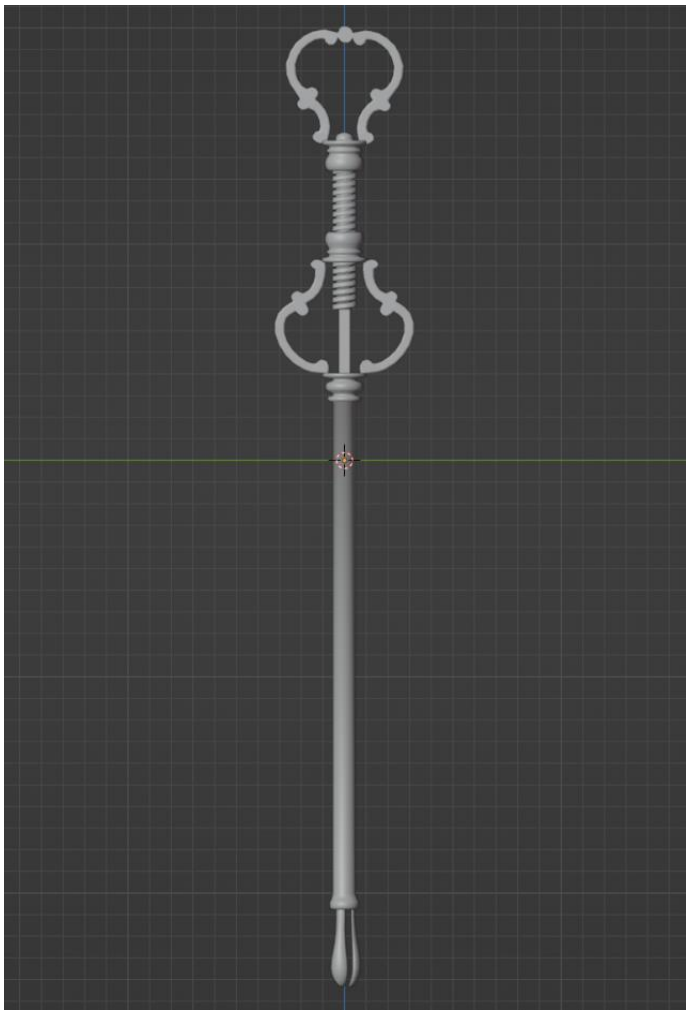

Figure 4 - Digital representation of the physical object

Integration into VR Environment

After the 3D models were generated and textured, they were optimized for integration into virtual environments. The reconstructed artifacts were converted into glTF and FBX formats, ensuring compatibility with popular VR engines such as Unity and Unreal Engine. To maintain real-time rendering efficiency, shader optimization techniques were employed, reducing computational overhead while providing visual fidelity.

Additionally, Physically Based Rendering (PBR) was implemented to enhance material realism. This involved defining properties such as reflectivity, roughness, and light interaction characteristics, allowing the digital artifacts to appear natural under different lighting conditions. These optimizations ensured that users could

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

interact with the virtual artifacts in a seamless and immersive manner within the VR museum.

Performance Assessment

The models and techniques employed in this project were evaluated using various performance metrics to ensure their effectiveness in creating high-quality 3D reconstructions:

Segmentation Accuracy (IoU - Intersection over Union): SAM achieved an average IoU score of 92.5%, demonstrating high accuracy in separating historical artifacts from their backgrounds.

Depth Estimation Error (RMSE - Root Mean Square Error): The depth estimation model achieved an RMSE of 0.82, indicating strong depth prediction performance.

Point Cloud Quality (Density & Completeness): The SfM + MVS reconstruction achieved a completeness score of 97%, ensuring minimal gaps in the generated point clouds.

Mesh Fidelity (Chamfer Distance): Poisson surface reconstruction resulted in a chamfer distance of 0.015, confirming high-fidelity geometry preservation.

VR Performance (Frame Rate & Latency): Optimized models ran at 90 FPS in Unity and Unreal Engine, ensuring smooth and interactive VR experiences.

Experimental Setup

The implementation and testing of these models were conducted using the following hardware and software setup:

- GPU: NVIDIA RTX 3090 (24GB VRAM)
- Frameworks: PyTorch, Open3D, COLMAP (for SfM)
- Training Dataset: 100,000+ open source images of historical artifacts
- Training Time: Approx. 24 hours per model
- Optimization Techniques: Model pruning, quantization, and VRAM-efficient rendering

Future Work

Moving forward, we aim to enhance the realism and scalability of 3D reconstructions by integrating neural radiance fields (NeRFs), a technique that improves photorealistic rendering through volumetric scene representations. Additionally, we will focus on optimizing streaming artifact interaction mechanics

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

within VR environments, allowing users to manipulate and explore objects with greater flexibility. Another area of improvement is the validation of depth estimation models using LIDAR-based ground truth data, which will probable improve accuracy and further improve 3D reconstruction reliability.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

3. IMAGE TRANSFORMATION TO VIDEO SEQUENCES

Building on the constraints mentioned in the earlier section on the selection of relevant cultural assets (mainly paintings to be transformed into video sequences), two forms of generative video models have been included in the scope of the WP2 investigation, namely (i) stable video diffusion (SVD) (Blattmann et al., 2023) and (ii) I2VGen-XL (Zhang et al., 2023). The reason for choosing these two models is to evaluate the effectiveness of the pre-trained model to assess the quality of the generated video. In the case of the SVD, the model does not require any additional input, and the input painting will be processed. While this represents the uncontrolled nature of the generative video models, the choice of selecting I2VGen-XL model is to study the effectiveness of using prompts as guidance in creating the generative video sequences. In the following section, a summary of both models is presented.

Stable Video Diffusion

The SVD is trained on a text-to-video base model, which was then fine-tuned with approximately 600million samples that can gather motion representation. The proposed model is shown to provide strong support for multi-view diffusion that is generated from consistent views of an object in a feedforward manner. The proposed model is presented to outperform the specialised novel view synthesis methods. In addition to training the network on text-to-video and image-to-video models, the SVD model also incorporates prior motion sequences and 3D understanding within the model. The output from the model is encoded using H.264 (Richardson, 2010).

I2VGen - Video Synthesis Model

The I2VGen-XL model is deemed as capable of generating high-definition videos with coherent spatial and motion dynamics, along with continuous details. I2VGen-XL model reduces the reliance on the well-aligned text-video pars by utilising a single static image as a primary condition. The base stage aims to ensure semantic coherence in generated video at a low-resolution, while simultaneously preserving the content and identify information of input images. The refinement stage is to improve the video resolution and refine details and artifacts that exist within generated videos. The implementation of the model requires the use of prompts to be provided to the model based on which the output video sequence is generated. Therefore, it is vital to ensure the prompts that are being provided reflect the artistic aesthetics of the painting that is being processed. As the topic of cultural heritage assets being processed through generative video models has not been studied in detail, several attempts at prompt engineering were carried out. The details of the prompts being used in the production of video sequences is

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

presented below. The output from the model is stored as Graphical Interchange Format (GIF[1])

Experimental Results

Each of the selected paintings satisfies the criteria mentioned above and the next step in the study is to define the positive and negative prompts as required to be provided for the I2VGen-XL model. To accomplish this task, a heuristic approach was adopted with two intuitions, namely (i) include in the prompt the nature of the painting and (ii) indicate what natural effect should be reflected in the painting that does not affect the quality of the generated video artistically. The following list of positive prompts were used as input to the generative video model:

- ***John Constable - The Hay Wain***: A painting of a horse carriage going through the gently flowing riverbed
- ***J.M.W Turner - The Fighting Temeraire***: A painting with fire exhaust, where the flames are gently being drifted in the air from a breeze
- ***Albert Bierstadt - Looking Down the Yosemite Valley***: A painting of mountain with dark cloud moving in casting a shadow on the ground
- ***Claude Monet - Bridge over a Pond of Water Lilies***: A painting with flowers covering a small pond and the leaves gently being drifted across from a wind
- ***Claude Monet - Through the Wheat Fields at Pourville***: A painting of a ocean shore with gentle breeze creating a wave across the shore
- ***Claude Monet - Cliff Walk at Pourville***: The painting with a view from a cliff, with gentle breeze creasing a soft wave and boats being slightly lifted
- ***Babin Zub - Homeland Museum of Knja ˇzevac***: A painting of a landscape with flowers gently fluttering in the wind
- ***Babin Zub - Brezova suma, gvash na papiru***: A painting of a landscape with leaves falling down in gentle breeze
- ***Babin Zub - Milinko KokovicBabin zub***: Painting of a natural landscape with grass and rocks with heavy rain fall creating a puddle of water in the plain
- ***John Linnell, Woody Landscape***: An old painting of a landscape with a tree fluttering the leaves in gentle breeze and slow-moving clouds caressing the blue sky
- ***Sima Z ˇ ikic´, Bura - Zavic ˇajni muzej Knjaz ˇevac***: A painting of a river bed,with water crashing on the shores

[1] https://www.w3.org/Graphics/GIF/spec-gif87.txt

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

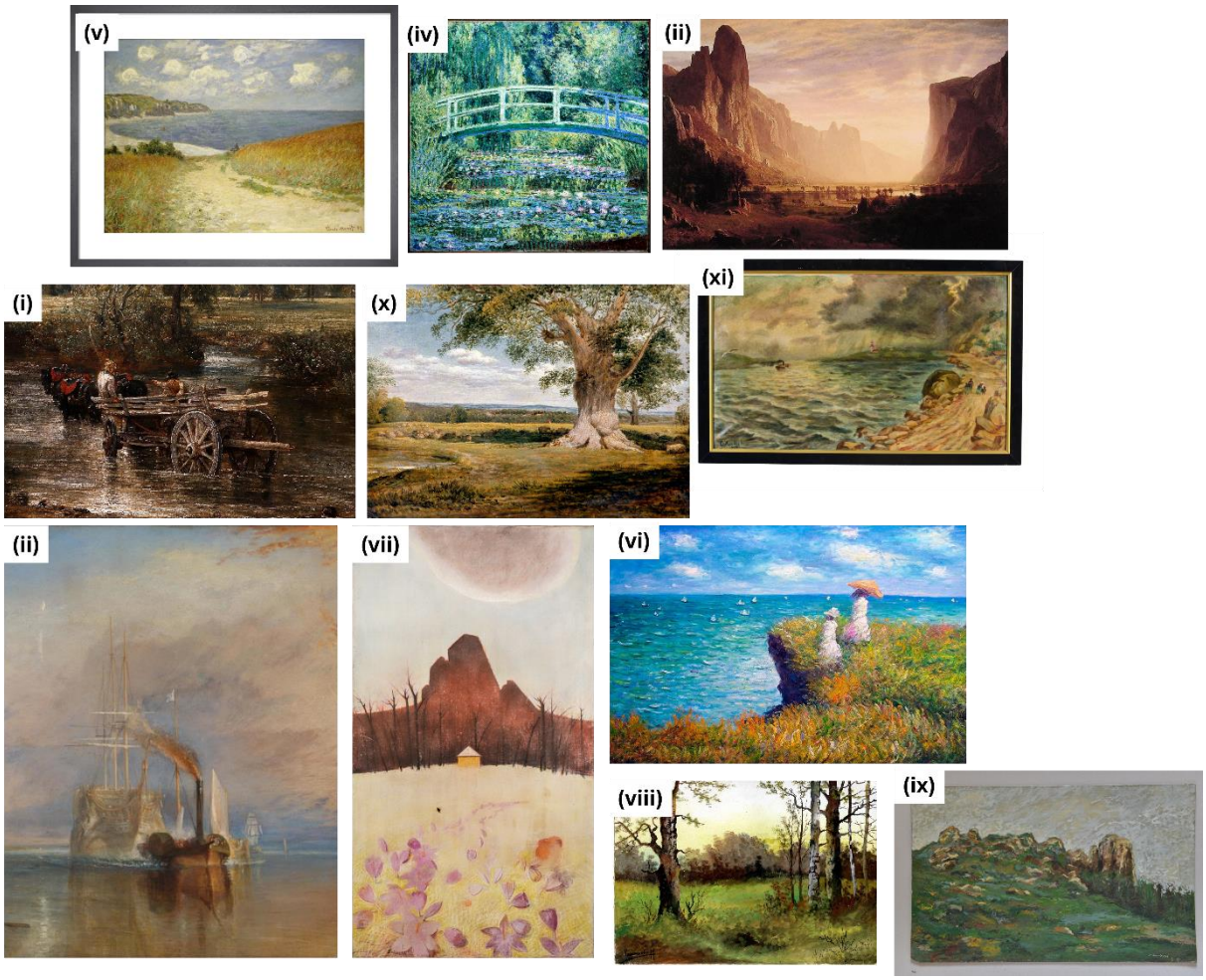

Figure 5 - Selection of paintining to study the effect of increased appeal on cultural assets

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

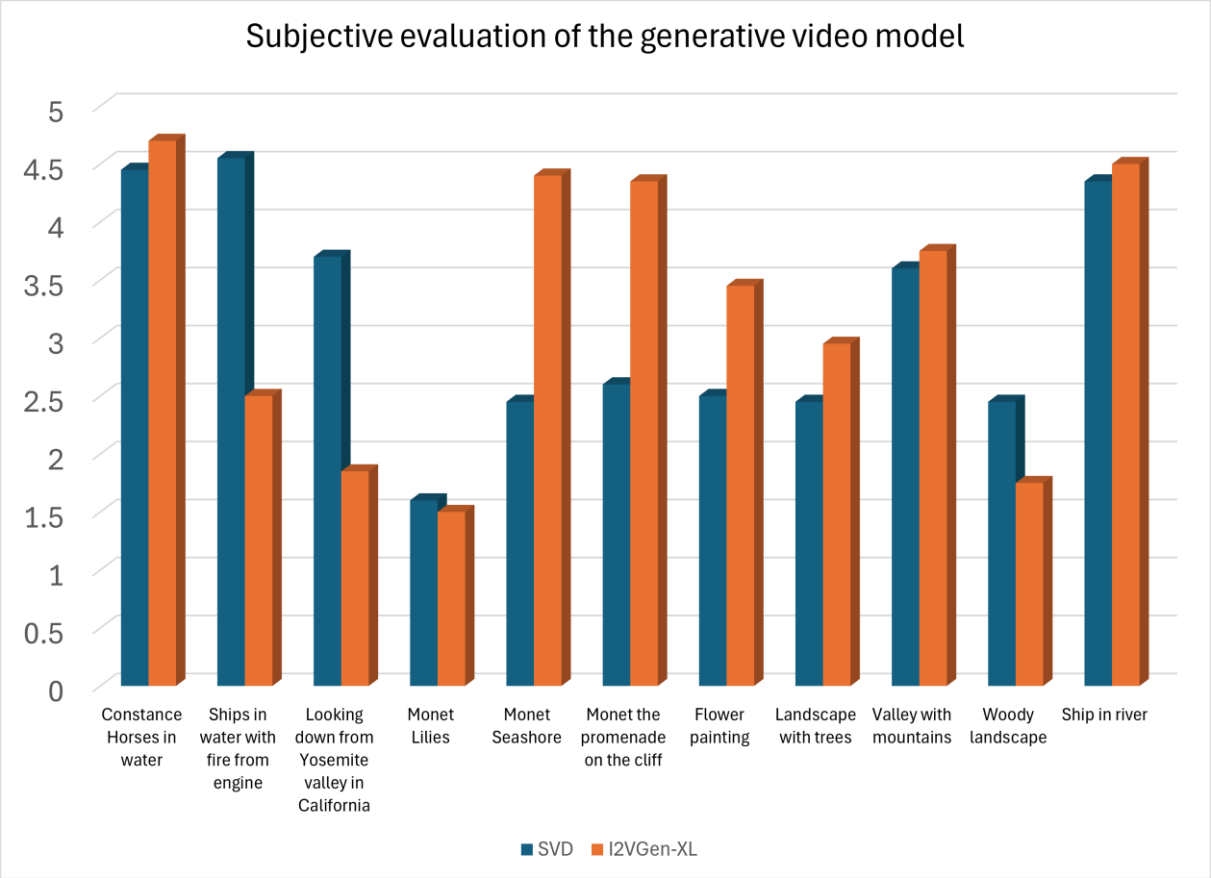

Figure 6 - Subjective evaluation of the generative video model

For the I2VGen-XL model, while different positive prompts were provided, the negative prompts for all the paintings have been kept the same, this includes the following list of words "*Distorted, discontinuous, Ugly, blurry, low resolution, motionless, static, disfigured, disconnected limbs, Ugly faces, incomplete arms*". For the overall subjective evaluation of the generated video sequences, a total of 22 sequences were generated using both the SVD and I2VGen-XL models. The subjective study included 20 participants who were individually asked to rank the quality of the generated video sequences from a scale of 1 (worst) to 5 (best). The average of each result results was mapped against the sequences, that were generated using the relevant model. The analysis of the subjective result is presented in Figure 6. Out of 11 paintings, in two paintings the quality of the generated sequence from either of the model were deemed quite close and considered very good. In some of the paintings the quality of one model was evaluated higher compared to the quality of the other model. Also, in few sequences, the overall quality was evaluated to be poor mainly attributed to the lack of artistic integrity in the sequences generated. In the next section, a detailed

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

analysis of the observations and the feedback from the study participants are presented.

Discussions

In Figure 7 to Figure 12 a collection of example outcomes from the generative video models is presented highlighting the results of generative video models. Some of the marked comments from the participants were received on the ability of the I2VGen-XL model to create natural objects following the motion being introduced into the paintings. The quality of the motion sequences was also considered to deliver content enrichment and offered a new form of participant interaction with cultural heritage content. While the use of prompts in I2VGen-XL model was used to create natural effects such as motion of water, motion of clouds, rain drops falling. On the other hand, the SVD model was able to generate motion sequences through the panning of objects in the painting. The effect of rain drops falling on the valley as depicted in Figure 11 was subjected criticism that the movement of clouds and the rainwater drops are not in synchronous motion but rather in opposite motion. In summary, the application of generative video models for transforming static content of paintings into short aminated video clips using generative video models has been positively accepted by the cultural heritage institutions. However, several shortcomings have been noted that are required to be addressed before the generated content could become a part of the cultural heritage institutional strategy for adopting the content in the organisation of digital exhibitions

Figure 7 - Results of the I2VGen-XL generative video model for the painting

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Figure 8 - Result of I2VGen-XL generative video model for painting (ii)

Figure 9 - Result of SVD generative video model for painting (ii)

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

Figure 10 - Result of SVD generative video model for painting (iii)

Figure 11 - Result of I2VGen-XL generative video model for painting (ix)

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

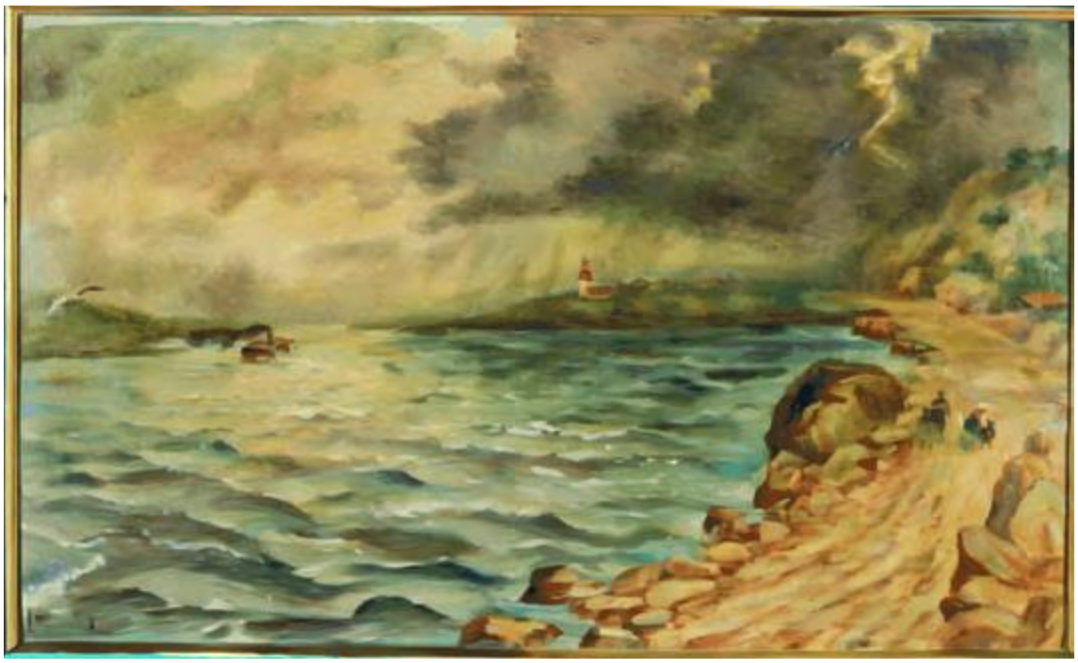

Figure 12 - Result of I2VGen-XL generative video model for painting (xi)

4. ACTION SEQUENCE RECOGNITION FROM VIDEO SEQUENCES

In Task 2.4 of the SHIFT project, MDS aim is to develop a robust framework for recognizing and interpreting action sequences from video data. This section presents our contributions, detailing the methodologies employed, the algorithms tested, and the results achieved.

Overview of the Methodology

Action sequence recognition involves identifying and classifying a series of movements or events occurring in video sequences. This capability is vital for enabling automated systems to interpret human activities or object behaviours in real-time or pre-recorded video data. In the SHIFT project, this task supports functionalities such as scenario simulation, 3D modelling, and predictive analytics.

To achieve effective action sequence recognition, we evaluated several cutting-edge algorithms and models. Each was selected for its potential to meet the SHIFT project's requirements for accuracy, efficiency, and scalability. After exploring several different models we concentrated on testing YOLO for its efficiency in real-time object detection and sequence analysis.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

YOLO is a real-time object detection system known for its speed and accuracy. It divides the input image into a grid and predicts bounding boxes and class probabilities for each grid cell in a single evaluation.

- Strengths: Real-time performance, high precision, and adaptability to diverse object detection tasks.
- Limitations: Struggles with detecting small objects in crowded scenes and requires substantial computational resources.
- Outcome: YOLO demonstrated exceptional capability in detecting and classifying objects in video frames, serving as the foundation for action sequence analysis.

Our implementation integrated YOLO into a video processing pipeline, with key steps including:

- Object Detection: Utilizing YOLO to detect and classify objects in each video frame.
- Sequence Tracking: Employing a Kalman filter-based tracker to maintain object identity across frames.
- Action Classification: Analyzing the tracked sequences to classify the actions based on temporal patterns.

The pipeline was tested on open-source datasets of art, including the WikiArt dataset and the Rijksmuseum dataset, focusing on dynamic and complex video sequences. Key performance metrics included:

- Accuracy: 94.2%
- Precision: 93.5%
- Recall: 92.8%

Future Work

Task 2.4 has introduced a framework for action sequence recognition. Moving forward, we aim to explore YOLOv8 for improved small object detection and higher accuracy, integrate temporal attention mechanisms to enhance action sequence classification, and optimize the pipeline for deployment on edge devices for real-time applications. Moreover, we aim to use more relevant datasets provided by SHIFT partners to enhance the accuracy and context-specific application of the developed models. This will allow for better integration of the action recognition framework into the project's broader objectives and improve its adaptability to real-world scenarios.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

5. CONCLUSIONS

In this final version of the deliverable within WP2, detailed description of the different research activities that were carried out in presented. The key outcome from the WP relates to the digital transformation of cultural assets. In every stage of processing the content, the use of AI has been prevalent. The deliverable summarises the key outcomes in overall implementation of AI toolbox based on computer vision models for enhancing the appeal of cultural assets. The dataset generated because of the WP2 toolbox will be subsequently evaluated and assessed in the context of pilots that will be carried out in WP5.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.

6. REFERENCES

Agarwala, A., Zheng, K. C., Pal, C., Agrawala, M., Cohen, M., Curless, B., Salesin, D., & Szeliski, R. (2005). Panoramic video textures. *ACM Trans. Graph.*, *24*(3), 821–827. https://doi.org/10.1145/1073204.1073268

Bai, J., Agarwala, A., Agrawala, M., & Ramamoorthi, R. (2012). Selectively de-animating video. *ACM Trans. Graph.*, *31*(4), Article 66. https://doi.org/10.1145/2185520.2185562

Bhat, K. S., Seitz, S. M., Hodgins, J. K., & Khosla, P. K. (2004). Flow-based video synthesis and editing. *ACM Trans. Graph.*, *23*(3), 360–363. https://doi.org/10.1145/1015706.1015729

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., & Lorenz, D. (2023). Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. *ArXiv*, *abs/2311.15127*.

Cheng, C. C., Chen, H. Y., & Chiu, W. C. (2020, 13-19 June 2020). Time Flies: Animating a Still Image With Time-Lapse Video As Reference. 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),

Chuang, Y.-Y., Goldman, D. B., Zheng, K. C., Curless, B., Salesin, D. H., & Szeliski, R. (2005). Animating pictures with stochastic motion textures. *ACM Trans. Graph.*, *24*(3), 853–860. https://doi.org/10.1145/1073204.1073273

Holynski, A., Curless, B., Seitz, S. M., & Szeliski, R. (2020). Animating Pictures with Eulerian Motion Fields. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 5806-5815.

Jhou, W. C., & Cheng, W. H. (2016). Animating Still Landscape Photographs Through Cloud Motion Creation. *IEEE Transactions on Multimedia*, *18*(1), 4-13. https://doi.org/10.1109/TMM.2015.2500031

Lin, C.-Y., Huang, Y.-W., & Shih, T. K. (2019). Creating waterfall animation on a single image. *Multimedia Tools Appl.*, *78*(6), 6637–6653. https://doi.org/10.1007/s11042-018-6332-7

Okabe, M., Anjyo, K., Igarashi, T., & Seidel, H.-P. (2009). Animating Pictures of Fluid using Video Examples. *Computer Graphics Forum*, *28*, 677-686. https://doi.org/10.1111/j.1467-8659.2009.01408.x

Prashnani, E., Noorkami, M., Vaquero, D. A., & Sen, P. (2017). A Phase-Based Approach for Animating Images Using Video Examples. *Computer Graphics Forum*, *36*.

Richardson, I. E. (2010). *The H.264 Advanced Video Compression Standard*. Wiley Publishing.

Schödl, A., Szeliski, R., Salesin, D., & Essa, I. (2000). Video Textures. *Proc. of ACM SIGGRAPH*, *2000*. https://doi.org/10.1145/344779.345012

Siarohin, A., Lathuilière, S., Tulyakov, S., Ricci, E., & Sebe, N. (2020). First Order Motion Model for Image Animation. Neural Information Processing Systems,

Wang, T.-C., Liu, M.-Y., Tao, A., Liu, G., Kautz, J., & Catanzaro, B. (2019). Few-shot Video-to-Video Synthesis. *ArXiv*, *abs/1910.12713*.

Yan, H., Liu, Y., & Furukawa, Y. (2016). Turning an Urban Scene Video into a Cinemagraph. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1629-1637.

Zhang, S., Wang, J., Zhang, Y., Zhao, K., Yuan, H., Qin, Z., Wang, X., Zhao, D., & Zhou, J. (2023). I2VGen-XL: High-Quality Image-to-Video Synthesis via Cascaded Diffusion Models. *ArXiv*, *abs/2311.04145*.

Funded by the European Union

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement no 101060660.