UNIVERSITY OF CAMPINAS

INSTITUTE OF COMPUTING

Master's Qualifying Exam

December 19th, 2024

Crafting Landscape Videos using Artificial Intelligence Models

Candidate: Felipe Santana Dias

Advisor: Prof. Dr. Zanoni Dias

Co-advisor: Prof. Dr. Hélio Pedrini

Contents

Abstract

Artificial Intelligence (AI) models are producing ever more realistic images. Yet, a core challenge persists: controlling this generation to meet the high standards and creative control of the entertainment industry in film, gaming, and digital art. Models such as Meta's Movie Video Gen and OpenAI's Sora, composed of transformers and diffusion methods, showcase remarkable visual outcomes — yet still lack control over specific elements, often leading to hallucinations. We address this challenge by combining object detection, natural language processing, and a suite of AI models to craft immersive and controlled videos. As outcome, we will create a visualization of a chosen landscape throughout the seasons, opening new creative frontiers and enabling the generation of virtual worlds and digital storytelling.

Chapter 1

Introduction

In this chapter, we define the research problem addressed in this work, the main motivations, the research questions, the planned objectives, and the structure of the text.

1.1 Motivation

We find ourselves at the intersection of physical reality and digitally constructed environments, where our increasing dependence on technology has evolved beyond functionality to provide emotional, communicative, and intellectual frameworks for the current generation [44]. This symbiotic relationship between technology, society, and the physical world forms the central focus of this research.

Technology's capacity to spread knowledge has granted society unprecedented access to information, dramatically expanding our horizons. However, this progress is not without challenges. The same algorithms designed to help us navigate the overwhelming volume of data are also contributing to the polarization of viewpoints. Social media, driven by dopamine-reward loops, distorts our perception of time and space, while forming relationships through screens may dilute human connection [10]. Although environmental awareness is growing, it often comes at the expense of these connections [13, 16]. These are critical issues that can be addressed not only through philosophical discourse but also by computational scientists developing new technologies.

Within this context, the concept of synthetic realities emerges—digital creations or augmentations facilitated by artificial intelligence methods. This concept highlights the efforts of both companies and researchers to integrate human experiences into the digital sphere [5]. The field of Digital Humanities (DH) further explores these possibilities, employing methods such as visualizations of extensive image collections, 3D modeling of historical artifacts, and alternate reality games [15].

Recent advances in hardware facilitating immersive experiences, exemplified by products such as Apple's Vision Pro[1], Meta Quest[2], and PlayStation VR2[3], along with innovations in chip and processor technologies integrating AI, are recalibrating user expectations of the technological landscape. Software is now poised to dynamically generate personalized experiences within these devices, marking a significant paradigm shift.

A study by Twitter reflects the growing demand for "Imaginative Escapism", where users seek

[1] `https://www.apple.com/apple-vision-pro`
[2] `https://about.meta.com/technologies/meta-quest`
[3] `https://www.playstation.com/ps-vr2`

immersive experiences that go beyond traditional gameplay and extend into augmented reality and physical collectibles [3]. This trend exemplifies how the line between physical and digital worlds is becoming increasingly blurred, driven by both user desires and advancing technology.

In the realm of visual interfaces, human-computer interaction design has historically evolved from one-dimensional representations into more complex 3D and multi-dimensional frameworks [36], facilitating deeper insights and understanding. Today, AI generative models such as DALL·E[4], MidJourney[5], Stable Diffusion[6], Kaiber[7], and Deforum[8] are pushing the boundaries of creativity. These models are progressing from generating still images to crafting short videos, moving towards creative autonomy.

As synthetic realities develop, they offer the potential to replicate and manipulate complex virtual environments, much like how Google Earth once mapped the physical world [43]. The future may hold the possibility of dynamically generating and altering digital versions of the Earth, providing new tools for visualizing the environmental impacts of human decisions—whether related to climate change or urban development.

This project embarks on a journey to explore the boundaries of synthetic realities, investigating the intersection of technology, creativity, and human experience. Through an interdisciplinary lens, we aim to uncover the transformative potential of AI-driven technologies in redefining our understanding of reality and charting new frontiers in creative expression.

1.2 Problem Description

Today, generative models face limitations in their ability to precisely control and specify the output they produce. Altering an object requiring changing its context often leads to distortion and loss of fidelity [8].

Nature and its environments possess an inherent, dynamic complexity; each day brings a unique visual composition, but also with similarities. Yet, within this variation, there are constraints governing the behavior and characteristics of individual elements such as the sky, clouds, trees, or buildings.

When employing AI to generate images of natural scenes, there's a lack of understanding regarding how each element interacts as a distinct entity within the environment. This limitation becomes evident when attempting tasks such as simulating a snowy day in Rio de Janeiro. The model may produce an image, but without a comprehensive grasp of how each element should respond, it often merges images of snow with the Rio de Janeiro scenery, resulting in a distorted representation (Figure 1.1).

The significance of addressing this challenge lies in its potential to create dynamic virtual worlds. Rather than static animations, these environments could evolve organically over time, mirroring the intricacies of nature. For instance, by incorporating weather patterns, we could observe how individual elements adapt and interact to shape the overall scene.

Essentially, this problem stems from the need to develop greater control over generative models by attributing context and semantics to the alteration of content previously created by the models, addressing the issues of overfitting data, hallucination, and consciousness.

[4] https://openai.com/product/dall-e-2
[5] https://www.midjourney.com
[6] https://stablediffusionweb.com
[7] https://kaiber.ai
[8] https://github.com/deforum-art/deforum-stable-diffusion

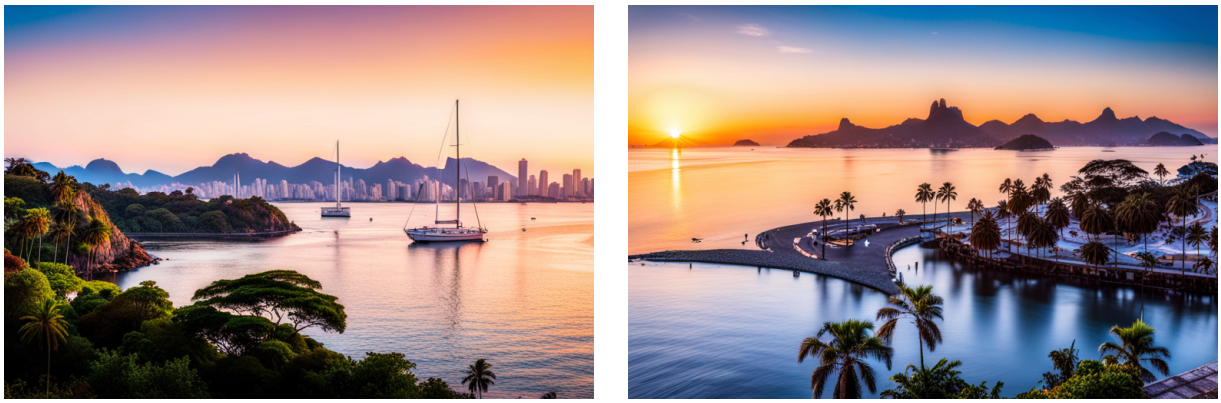

Figure 1.1: Rio de Janeiro landscape generated by StableDiffusion using the following prompt modifications: (left) clear sky, lush vegetation, few boats in the water, calm sea; (right) clear sky, vegetation covered in snow, frozen water, sunset lights, calm sea.

1.3 Research Questions

In the domain of Artificial Intelligence (AI), the manipulation of visual imagery is characterized by the fusion and composition of various similar images, allowing the generation of entirely new content. This phenomenon, known as hallucination, presents a unique opportunity to explore creativity within AI systems. However, it also bring question on the degree of control we possess over the generated visual output. The following research questions will guide this project:

RQ1: How can we enhance our understanding of the mechanisms underlying AI's hallucinatory processes in visual content creation?

RQ2: What methodologies and techniques can be employed to exert greater control over the specific visual elements that AI hallucinates and reproduces?

RQ3: How does leveraging additional data sources beyond the prompt contribute to enhancing the precision and accuracy of AI-generated visual content by providing contextual and semantic information?

1.4 Objectives

Our objective for this project is to develop an immersive experience that dynamically generates and reproduces landscapes from video recordings under various weather conditions over time. This process must be heavily supported by AI models and/or a pipeline capable of generating realistic and cohesive videos, significantly speeding up the animation process by:

- Utilizing AI to manage and adjust generated content for a cohesive representation of reality.
- Enabling dynamic alterations of the landscape's perspective based on known variables.
- Transforming 2D photographs into realistic 3D animations.

Our goal is to push the boundaries of current AI capabilities by enhancing the artistic potential of generative models, moving from still images to 3D animations where we exercise greater control over visual elements beyond the initial input, allowing each element to exhibit its own behavior.

The potential impact of this research within the entertainment industry is, particularly, in expediting content production for games, movie scenarios, environmental effects, and artistic pieces.

With this research, we aim to create an immersive experience that can be accessed through video or augmented reality devices. On the academic front, we intend to produce articles on each component used and develop a "plug and play" pipeline that can evolve alongside advancements in AI technology, growing alongside the rapid advancements in AI.

1.5 Text Organization

This text is organized as follows. Chapter 1 describes the research problem to be addressed in this project along with the main objectives and contributions. Chapter 2 provides a description of essential concepts for understanding the proposed research and presents existing related works in the literature, as well as an initial bibliographic review. Chapter 3 describes the proposed methodology, the databases to be used, and the metrics used for model evaluation. Finally, Chapter 4 presents an action plan and a schedule for the execution of tasks related to the project. Finally, Chapter 5 presents some concluding remarks.

Chapter 2

Background

As the demand for immersive and dynamic environments grows, various industries have developed innovative approaches to tackle the challenges of creating complex visual scenes. This chapter explores how these challenges are currently being addressed in two key areas: cinema and games.

2.1 Cinema

Pixar, the renowned animation studio for its cutting-edge technology, has consistently pushed the boundaries of visual storytelling in the film industry. Known for its complex and dynamic imaginary worlds, Pixar's films often require the creation of the same environment across different contexts. "Elemental" [39], one of Pixar's releases in 2023, exemplifies the studio's advancements in creating intricate backgrounds.

In the elemental city, Pixar's pipeline introduces variation by shading instanced buildings through an image-based color picking approach, where the palette of the city is generated from one or more input images. The tool developed for the reassignment of local colors enables the application of palette shading to any assets, or the blurring of local colors across models, while preserving the material properties established during the shading process. This technique has proven effective in simplifying visual detail in specific frame areas [45].

Furthermore, Pixar has devised a system for art-directing the city skyline's silhouette, allowing artists to create 2D curves that are mapped onto different sections of the city, causing the corresponding buildings to shift vertically to align with the silhouette curves [45].

This approach offers two strong points: (i) the ability to change colors dynamically allows for environmental adjustments during different lighting conditions, making it ideal for daily scene transitions, and (ii) there is no need to rebuild any elements, streamlining the process.

The natural entropy and randomness of the image-based approach contribute to the final visual outcome, enhancing realism. It is primarily applicable to highly detailed, constructed elements with specific shapes and material properties. This requires a complex metadata structure to be effective.

Pixar's innovations are impressive and it is noteworthy that the company has shown resistance to incorporating AI into their creative processes. This is likely due to concerns about losing control over the generated images, as discussed in recent industry analyses [21].

2.2 Games

The New Campo Marzio project [31] is an interactive, AI-driven simulation that reimagines Piranesi's iconic architectural plan using advanced technology. The city is dynamically generated in real time as users explore it, with its form shaped by artificial neural networks (ANNs) and symbolic intelligent agents. The simulation uses Generative Adversarial Networks (GANs) trained on Piranesi's documents to produce two types of resolution: a low-resolution backbone for foundational navigation and a higher resolution for detailed visual recreation.

In the workflow, a lower-resolution map is first generated, allowing the AI to understand the spatial voids and characteristics of the environment. This forms a procedural navigation field composed of colliders, walkable surfaces, and evolving destinations. This field is crucial as it governs the interactions between user characters and the city's architecture.

Once the basic map is established, higher-resolution GANs refine the city's visual details, ensuring the environment is cohesive. The final output is a dynamic, ever-evolving version of Piranesi's Campo Marzio, merging modern AI techniques with historical architecture, creating an immersive and fluid user experience. This project explores new methods of architectural representation, blending AI-driven storytelling with game-engine technologies to create a unique, open-ended narrative world.

2.3 Technology

At the forefront of technological development, significant investments are being made by both tech giants and startups. On one side, we have commercial black-box models as Runway Gen3[1], LumaLabs[2], Kling1.5[3] that offer APIs for their use. On the other hand, access to academic advancements is crucial for understanding and leveraging foundational models in this context, as it helps us grasp the architectures that are progressively enabling greater control over image generation. Understanding the current capabilities of these models allows us to choose the most suitable ones to optimize our results. In this section, we will discuss two models: Movie Gen Video by Meta [28] and OpenAI Sora from OpenAI [25].

2.3.1 Movie Gen Video

Meta's product Movie Gen Video, is described as the "research for the most advanced media foundation AI models"[4]. This technology enables the generation and editing of videos up to 16 seconds using AI. Meta's access to extensive computational power and datasets allows it to run this 30 billion parameters model [28], and the architecture behind their results is of particular interest for study and consideration.

Image and video generation were carried out in a spatio-temporally compressed latent space, built using a Temporal Autoencoder model (TAE) based on a Variational Autoencoder [14,33]. The initial prompt was encoded to generate an embedding, which, along with noise, served as the conditioning for the model. This conditioning was fed into the generative model, and after processing, the output was decoded using the TAE to return results in the RGB spectrum. The architecture is presented in Figure 2.1.

[1]Runway Gen-3: `https://runwayml.com/research/introducing-gen-3-alpha`
[2]LumaLabs: `https://lumalabs.ai/dream-machine`
[3]KlingAI: `https://klingai.com/`
[4]Movie Gen: `https://ai.meta.com/research/movie-gen/`

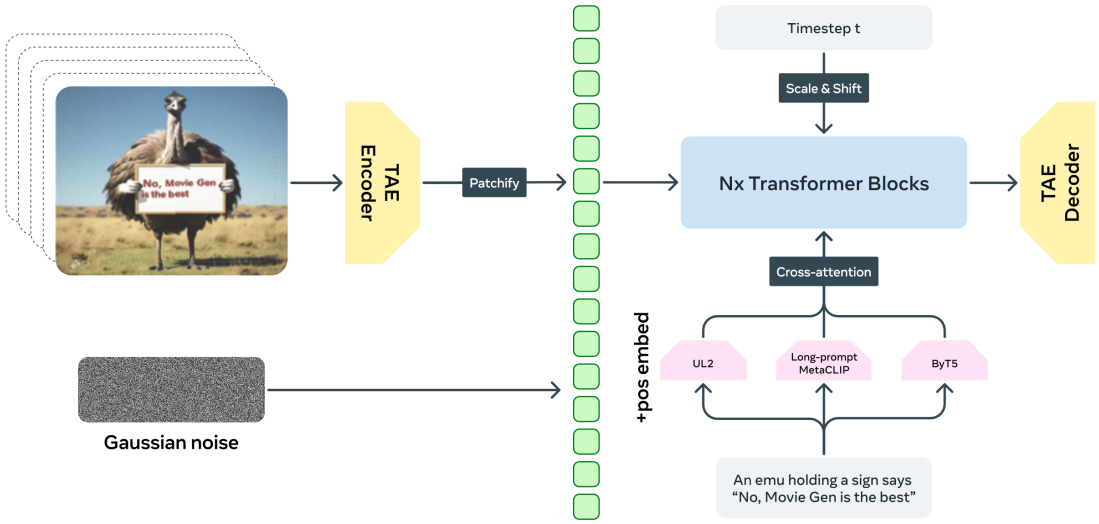

Figure 2.1: Overview of the joint image and video generation pipeline.

The transformer backbone follows the LLaMa3 architecture [19], utilizing cross-attention from the prompt generated embeddings using a combination of three pre-trained text encoders: UL2 [42], ByT5 [49], and Long-prompt MetaCLIP [48].

Given the challenges associated with the volume of information and the availability of large datasets, the training process for video generation was divided into phases. It began with single images and then incorporated time complexity by combining different frames as individual images. The editing capabilities were trained using a dataset that combines generated images with their corresponding prompts. As highlighted in the paper, the choice of a simplified architecture facilitates the scaling of model size and training.

Finally, for video generation from the individual images, they implemented tiled inference using the TAE, overlapping optimal frames during the reconstruction process (Figure 2.2).

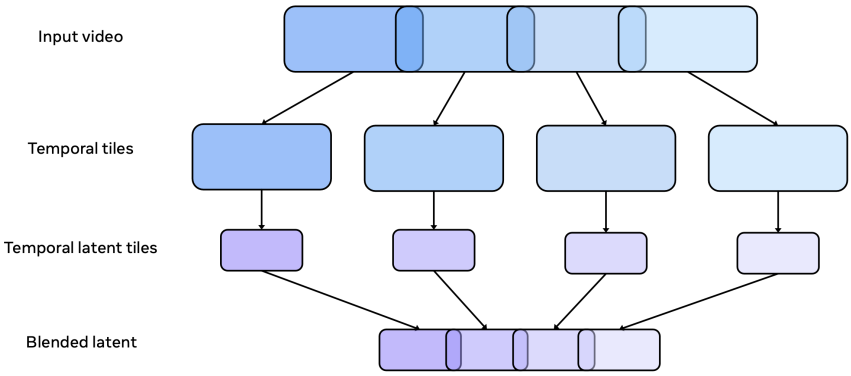

Figure 2.2: Tiled inference using the TAE.

2.3.2 OpenAI Sora

Sora is the AI model from OpenAI to create scenes from text instructions[5]. Although lacking an official paper detailing the model's architecture, it has a technical report that provides insight into its methods, capabilities, and limitations [25].

Sora adopts a diffusion transformer over GANs due to the scalability and performance benefits associated with these models [7, 27], aiming to reduce complexity while improving model training efficiency. Their results demonstrate that scaling up training significantly improves the model's generative capacity, achieving outstanding outputs, representing a promising path toward creating realistic simulations of physical environments without specific targeted training [25].

Sora processes "spacetime patches" as input, similar to the way current large language models (LLMs) operate. This approach unifies various forms of visual data, enabling efficient large-scale training. The dataset was developed by applying a re-captioning technique, introduced in DALL·E 3, to video data, producing detailed text captions for test set videos. Training with highly descriptive video captions was found to improve text fidelity and the overall quality of generated videos [2].

OpenAI has also published several studies on improving diffusion models, which may offer valuable techniques for further advancing generative capabilities. These include Consistency Models [20, 40] and Point-E [23] for generating 3D objects.

[5]OpenAI Sora: `https://openai.com/index/sora/`

Chapter 3

Methodology

Given that generative models development requires massive computational workforce, the workflow for this project aims to integrate the different technologies available in the market and literature, advancing as the models are publicly available. Due this nature of the field, it is essential to structure the workflow allowing modularity and ensuring that we can easily incorporate advancements.

The key stages of the workflow (Figure 3.1) include:

1. **Object Detection:** From an initial image/video, we identify the elements and context of the scenario.
2. **Prompt Generation:** For each element, a prompt is generated using large language models (LLM) to describe the object visual and behaviorly.
3. **Image Generator:** The created prompt will be used to generate element's images in the proposed new context.

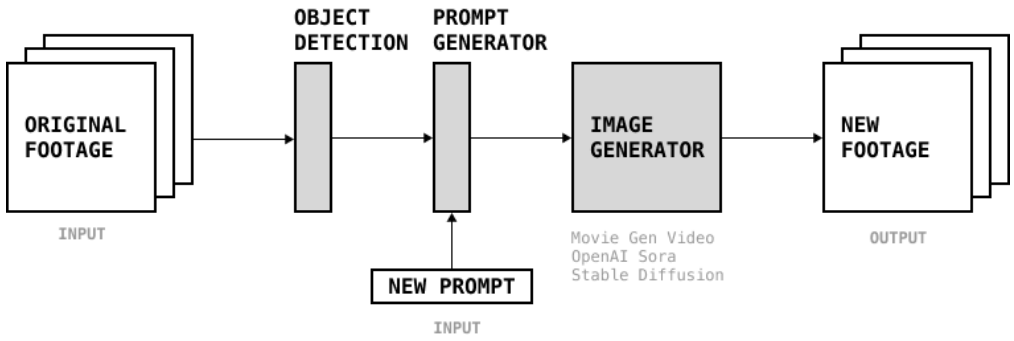

Figure 3.1: AI pipeline for enhancing footage: From original footage to generating new video using object detection, engineered prompts, and AI Image Generators.

We start by detecting and analyzing the different elements of an image. From there, prompts are automatically created to input the right conditions in the most advanced models available aiming to generate a consistent and realistic output.

Each of these steps will be composed of their own architecture and explore different technologies that will be interconnected. We will carry out an extensive research on state-of-the-art methods for

each component, ensuring a robust understanding of the parameters and their potential for synergy in the final creation process.

Although this level of granular control requires significant computational resources, it provides the flexibility for post-generation adjustments, giving artists greater creative power over the final product.

3.1 Object Detection

Based on recent surveys over object detection [6, 17, 53], we can notice that YOLO (You Only Look Once) is a popular algorithm that stands out for its speed and efficiency [30]. Unlike traditional methods that repurpose classifiers for detection tasks, YOLO uses a single convolutional neural network (CNN) to predict both bounding boxes and class probabilities directly from the input image in a single pass. This enables real-time detection, making it suitable for applications requiring quick processing, such as autonomous vehicles and video surveillance, which might be interesting when proposing real time transformation.

However, YOLO has its limitations when detecting small, closely grouped objects and objects of uncommon shapes or proportions since the model's loss function treats all bounding boxes similarly, which can affect accuracy for smaller objects.

For the current purpose of this project, it seems to be a better approach. A model generates region proposals using Region Proposal Network (RPN), which are then classified and refined by Fast R-CNN, in a unified framework, providing a balance between speed and accuracy [32].

Faster R-CNN provides high detection accuracy, making it suitable for scenarios requiring precise object recognition, which is the case of natural landscapes. RPN generate proposals quickly, improving overall processing speed and the multi-scale anchor boxes enable the model to handle objects of varying sizes and shapes, which is useful in complex environments.

The Faster R-CNN's pipeline consists of the following steps:

1. **Feature Extraction:** The input image is passed through a convolutional neural network (typically a pre-trained network such as VGG [37] or ResNet [11]) to extract a feature map. This map captures high-level information about the image, including edges, textures, and patterns.

2. **Region Proposal Generation:** The RPN takes the feature map as input and generates a set of candidate regions (or "proposals") that potentially contain objects. The RPN uses anchors, which are predefined boxes of different scales and aspect ratios, to generate these proposals. It then refines the positions and sizes of these anchors to better fit the detected regions.

3. **Region of Interest Pooling:** The proposed regions are then processed by the Region of Interest (RoI) pooling layer, which converts the varying sizes of proposals into fixed-size feature maps. This step allows the model to handle objects of different sizes while maintaining computational efficiency.

4. **Object Classification and Bounding Box Regression:** Each region is classified to determine whether it contains an object and, if so, what class of object it is. Additionally, the model refines the bounding box coordinates to improve the accuracy of object localization.

5. **Final Detection:** Non-maximum suppression is applied to eliminate redundant overlapping boxes and retain only the most accurate ones, producing the final object detection results.

By leveraging these features, Faster R-CNN can be effectively utilized for various applications, such as tracking animal populations, monitoring vegetation growth, detecting invasive species, and supporting conservation efforts through automated image and video analysis in diverse natural habitats.

In this project, from a single frame of a natural landscape, Faster R-CNN will generate multiple images from the element that will be used later on the behavioral analysis of each object (Figure 3.2).

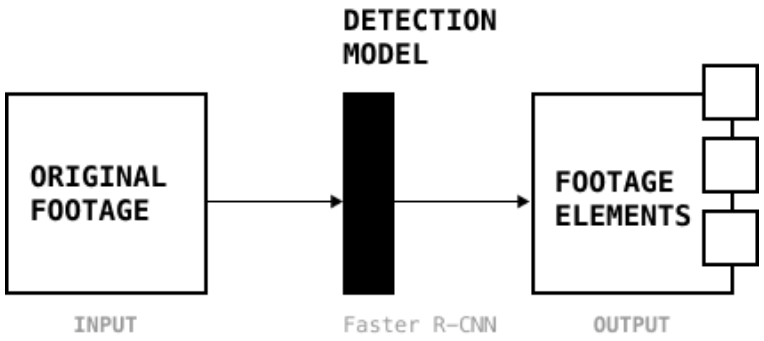

Figure 3.2: Pipeline for the object detection: From a single frame footage, Faster R-CNN will be used to generate images of the scenario elements.

3.2 Large Language Model (LLM)

In the creation of multiple images across various contexts, the ability to generate high-quality visuals relies heavily on two aspects: (i) precise object descriptions using Natural Language Processing (NLP) techniques, and (ii) prompt engineering to effectively integrate these objects into the target image.

3.2.1 Object Descriptions with NLP

In the field of NLP for image-to-text generation, several advanced technologies have emerged, each offering unique approaches to crafting meaningful textual descriptions from visual inputs. Some of the most notable techniques include [12]:

- **End-to-End Framework:** This approach utilizes an encoder-decoder model, where a Convolutional Neural Network (CNN) encodes an image into a visual feature vector. A Recurrent Neural Network (RNN), such as Long Short-Term Memory (LSTM) or Gated Recurrent Unit (GRU), then generates text based on that vector. This method efficiently links visual inputs with textual outputs, streamlining the process.
- **Attention Mechanism:** This method enhances the model's focus on specific regions of an image, which improves the relevance and accuracy of captions. By dynamically attending to different parts of an image during the text generation process, this technique is especially useful for images containing multiple focal points or complex scenes.

- **Compositional Framework:** This approach enables separate training for various components of the system, offering flexibility and scalability. By detecting key visual concepts, it combines image and text information to produce meaningful descriptions. This framework is more adaptable but may require specific fine-tuning for certain tasks.

The Attention Mechanism has proven particularly effective in generating high-quality, contextually accurate captions. It addresses some of the limitations of traditional end-to-end frameworks by allowing the model to adjust its focus dynamically based on the content already generated. This leads to more accurate, detailed captions, especially when dealing with complex or multi-object images [9, 12].

The process involves three key steps:

1. **Image Encoding:** A deep CNN encodes the input image into a global visual feature vector, capturing its overall semantic information.

2. **Attention-Based Focus:** During text generation, the model focuses on specific image regions, creating subregion visual vectors for the areas most relevant to the description.

3. **Text Generation:** The global visual feature vector, along with attention-modulated subregion vectors, is processed by an RNN-based decoder (e.g., LSTM or GRU) to generate the corresponding text, guided by both the overall image context and specific highlighted details.

In our context, applying the Attention Mechanism after the object detection can be experimented in different window sizes, from the exclusive object to bigger windows that give more context on the environment.

3.2.2 Prompt Engineering

The achieving desired results hinges on the ability to accurately describe the object [18, 26]. Several guidelines are proposed for the methodological production of high-quality images, as the following template [26]:

[Medium] [Subject] [Artist(s)] [Details] [Image repository support]

In our previous research, to create prompts for each element, we combine variables with preset words, ensuring consistent and high-quality results [8]. The object descriptions generated from the original input are merged with these preset elements, producing the final prompt used to generate the image (Figure 3.3).

For this project the prompt generator will use LLMs to describe the current image, generating prompts that will work as input in AI Models as Stable Diffusion. The content generate will be compared with the original one by similarity, penalizing the ones with greater difference from the original image.

The descriptions generated of the images will be stored as database for futures implementations of the project.

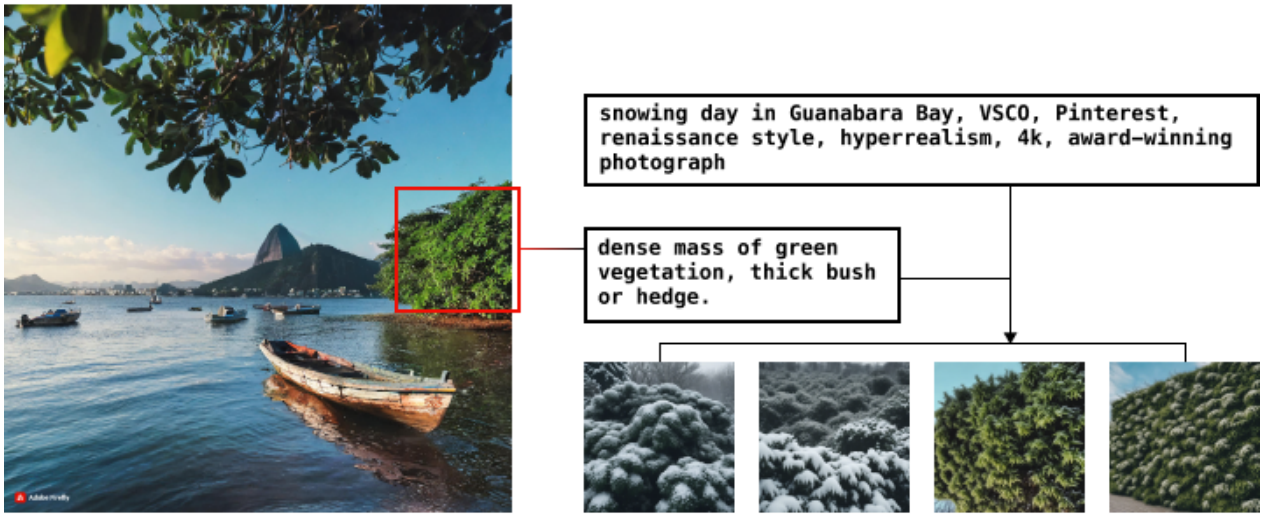

Figure 3.3: Visual reinterpretation of Guanabara Bay with AI-generated styles: A snowy scene.

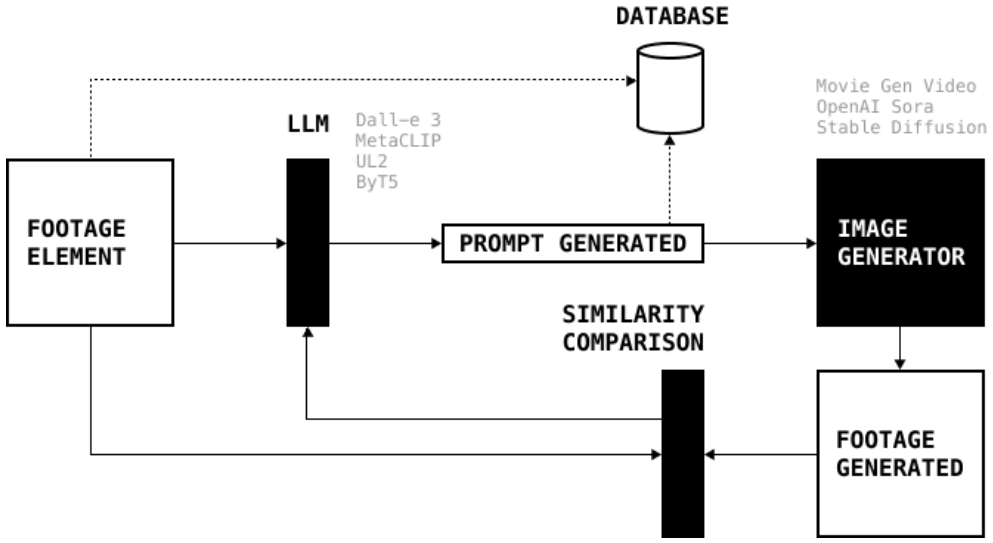

Figure 3.4: Pipeline for the prompt generation: From a single element, a description is made and evaluated by its final result in a generative model.

3.3 Generative Adversarial Networks

In this study, we aim to experiment with image generation using tools as Stable Diffusion API, Adobe Firefly[1], and DALL·E, leveraging both previously generated images and entirely new prompts. This will help us explore how varying parameters affect the maintenance of context in the generated images. By doing so, we will gain a deeper understanding of how different techniques and inputs influence the outcome.

Additionally, we aim to leverage a pre-trained model on a large-scale image database, further

[1]Adobe Firefly: `https://www.adobe.com/products/firefly.html`

training it on our own dataset of images and generated prompts. This dataset will integrate metadata, including weather conditions, geographic coordinates, and other measurement data. In alignment with the work by Meta [28], we will employ a Temporal Autoencoder Model (TAE) to encode this multimodal data into a unified latent space. Our goal is to generate images within this latent space, allowing us to reconstruct them having the metadata as input parameters (Figure 3.5).

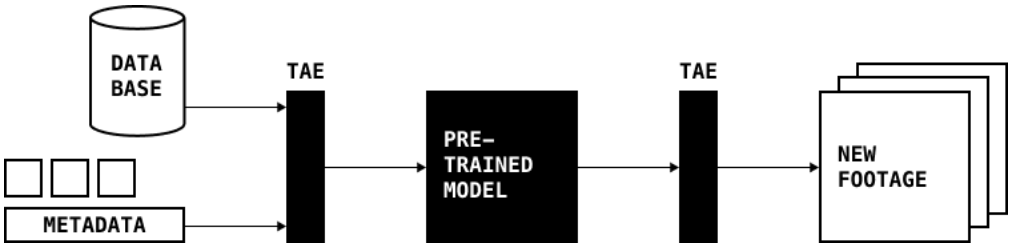

Figure 3.5: Pipeline for the generative model: Input is encoded by a Temporal Autoencoder Model to train the model in the latent space.

3.4 Inpainting Models

Inpainting is a technique used in image processing to restore missing or damaged parts of an image consistent and visually plausible with the surrounding content [51]. In image restoration, it helps repair damaged photographs by filling in missing areas of old or deteriorated images with textures and colors that blend seamlessly with the original. For object removal, it is used to eliminate unwanted elements or blemishes while preserving the visual coherence of the surrounding area. In augmented reality, inpainting enhances digital overlays on real-world images by filling in background spaces that become visible when objects are moved or removed. It is also widely used in video editing to ensure smooth visual continuity by filling gaps left when objects or elements are removed from video sequences.

The inpainting landscape is characterized by a variety of technologies:

- **Structural Information-Guided Inpainting:** Utilizes structural information from known image regions (such as edges or segmentation maps) to guide the inpainting process. This approach ensures structural coherence in the restored areas, particularly useful in images with sharp details or significant structural elements [46].
- **Attention-Based Inpainting:** Leverages attention mechanisms to focus on relevant parts of the image, computing similarities between image patches to borrow or copy information from distant spatial locations. This method effectively handles complex scenes by reconstructing textures and structures in large missing regions [46].
- **EdgeConnect:** Integrates adversarial edge learning into its inpainting process, using a two-stage approach of edge detection and image completion. This method generates a rough edge in the missing areas to serve as a priori information for the image completion network, which refines the details based on this structure [29].

Attention-Based Inpainting is considered one of the most advanced techniques due to its ability to borrow or copy information from distant parts of an image. This allows for more precise and contextually appropriate results, especially when dealing with large missing areas or complex scenes.

The process starts with image encoding, where a deep neural network, often a convolutional neural network (CNN), transforms the input image into a global visual feature vector. The attention mechanism then focuses on specific areas of the image relevant to the inpainting task. By comparing image patches, the model can borrow details from other parts of the image to fill in the gaps, ensuring the result blends naturally with the rest of the image [46].

Another cutting-edge approach, EdgeConnect, is known for its innovative use of adversarial edge learning. This method stands out for its ability to handle structural information effectively, delivering high-quality inpainting results. EdgeConnect's process begins with edge detection, where the edges of the image are identified, including rough outlines for missing areas. This edge map then guides the image completion network, using the structural information to fill in and refine the missing sections, resulting in a realistic and cohesive final image [29].

By combining attention-based inpainting with EdgeConnect, we can create images that look natural and maintain consistency across all elements. These advanced techniques allow for detailed adjustments to individual parts of the scene while ensuring the overall composition remains visually coherent, enhancing the realism of the final product.

3.5 Datasets

Our primary training efforts will focus on constructing the large language model (LLM), where we will compare the image descriptions, the generated prompts, and the resulting images from various AI models. We will initiate training on image datasets of landscapes available in public repositories found on Kaggle [34, 35, 38]. Later in the study, we will extend this approach to video datasets, following a similar methodology [47].

For the generative models, given the current market availability and capacity for large-scale training, we will pursue two strategies. First, we will evaluate pre-trained models, such as VGG[2] and ResNet[3] on extensive datasets, which offer robust baseline performance with manageable computational demands.

Secondly, we will access the Large Nature Model[4], which provides specialized capabilities for training on nature datasets. However, this model is computationally intensive, and thus, its use will depend on the project's scale and resource constraints.

3.6 Evaluation Metrics

To assess the visual fidelity of synthetic realities generated by AI models, metrics such as Structural Similarity Index (SSIM) [1,4] provide quantitative measures of how closely generated images resemble real-world references in terms of structure, detail, and noise levels. These metrics are interesting when changing visual elements or restoring parts of an image, ensuring that the modifications maintain the integrity of the original scene. Another important measure, Fréchet Inception Distance (FID) [1], is particularly useful for evaluating generative models, as it compares the distribution of generated images to real ones, offering a score that reflects the realism of the synthetic content.

To ensure that generated images remain cohesive and contextually appropriate, Content Consistency and Context metrics are necessary. These include contextual loss [22], which evaluates how well generated elements fit into their surroundings without disrupting the overall coherence of the

[2]VGG: `https://github.com/pytorch/vision/blob/main/torchvision/models/vgg.py`

[3]ResNet: `https://github.com/pytorch/vision/blob/main/torchvision/models/resnet.py`

[4]Large Nature Model: `https://refikanadol.com/works/large-nature-model-living-art/`

scene. In scenarios where attention-based models are used, metrics such as attention-weighted semantic alignment [52] can help measure whether the AI model has properly focused on key objects or regions in the image. Additionally, edge preservation metrics [50] ensure that structural elements, such as the boundaries between objects, remain sharp and intact, particularly during processes such as inpainting, where the AI fills in missing parts of an image.

When expanding synthetic realities into the realm of animation and 3D environments, Temporal Metrics become vital for evaluating the fluidity and realism of the generated scenes [24, 41]. Frame consistency, for example, measures how smoothly transitions occur between frames in an animation, ensuring that the motion feels natural rather than jarring. This is particularly important when simulating dynamic elements such as weather changes or environmental effects. Motion Smoothness is another key metric, assessing the fluidity of object movements or transitions, which contributes to the overall immersion and believability of the animated or 3D environment. These temporal metrics ensure that not only individual frames are realistic, but that the sequence as a whole maintains coherence over time.

Chapter 4

Project Timeline

Our objective is to deliver value at each stage of development, progressing through key phases from static images to fully immersive 3D environments. We will adopt an MVP (Minimum Viable Product) approach, starting with image construction, transitioning to animated images, and ultimately creating dynamic 3D environments (Table 4.1).

4.1 Phase I – Static Images

In the first phase, we will generate static images with altered environmental conditions. For example, starting with an image of Rio de Janeiro, we aim to transform it into a snow-covered landscape, moving beyond realism into the hallucinatory spectrum that AI-generated imagery can achieve. This phase will focus on adjusting individual elements such as trees and the ocean, where comparable visual examples exist in different weather conditions. The goal is to explore how these altered elements can still maintain the integrity of the overall landscape.

- **Estimated Duration:** 3 months
- **Output:** The same landscape represented under different weather conditions.

4.2 Phase II – Animated Images

In the second phase, we will introduce movement by adding a temporal dimension to the static images. Starting with a calm landscape video, we will simulate a dynamic weather change, such as transitioning from a peaceful day to a stormy scene. The key challenge here is creating smooth, controlled transitions for each element, such as wind effects on trees or the evolving state of the ocean.

- **Estimated Duration:** 3 months
- **Output:** Animations showing different weather conditions within the same landscape.

4.3 Phase III – 3D Environments

The final phase will involve adding spatial dimensions to the animated environments. We aim to create videos that move beyond the confines of the original footage, expanding perspectives without

Table 4.1: Project's timeline segmented by bimonthly.

ACTIVITIES	**1st Year**						**2nd Year**					
	1	2	3	4	5	6	1	2	3	4	5	6
PREPARATION	•	•	•	•	•	•						
Mandatory credits in subjects.					•	•						
Bibliographic review	•	•	•	•	•	•	•	•	•	•	•	
EQM			•	•								
PHASE I: Static Images							•	•				
Object Detection Model							•					
Prompt Generation							•					
Image Generation							•	•				
Inpainting Model								•				
PHASE II: Animated Images								•	•			
Temporal Generation								•	•			
PHASE III: 3D Environments										•	•	•
Spatial Generation										•	•	•
CONCLUSION											•	•
Document and publish results											•	
Writing the dissertation document											•	•
Defense of Master's Thesis												•

compromising the context or realistic appearance of the content. By doing so, we can explore the landscape from multiple angles, generating a more immersive experience.

- **Estimated Duration:** 6 months
- **Output:** A fully rendered video that moves fluidly through the environment, providing multiple perspectives.

This phased approach will allow us to iteratively test and refine our methods, building upon the technology stack while delivering meaningful outputs at each stage.

Chapter 5

Conclusions

This study presents a framework for enhancing the creation of synthetic realities through AI-powered generative models. The integration of object detection, natural language processing, and GANs has demonstrated the potential to move beyond static images and simple animations to more complex, dynamic environments. By developing a pipeline that allows for real-time modifications to generated landscapes and maintaining contextual coherence, we have expanded the possibilities for immersive experiences in virtual environments.

This research will also contributes to a deeper understanding of how AI can augment creative processes, offering new tools for industries such as gaming, film, and environmental modeling. Future developments in AI technology will undoubtedly push these boundaries further, and the modular nature of our workflow ensures that this system can evolve alongside these advancements

Bibliography

[1] S. S. Baraheem, T.-N. Le, and T. V. Nguyen. Image synthesis: a review of methods, datasets, evaluation metrics, and future outlook. *Artificial Intelligence Review*, 56(10):10813–10865, 2023. 15

[2] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, W. Manassra, P. Dhariwal, C. Chu, and Y. Jiao. Improving image generation with better captions. `https://cdn.openai.com/papers/dall-e-3.pdf`, 2023. Accessed: 2024-10-31. 8

[3] Black Swan Data, CrowdDNA, commissioned by Twitter. The Conversation: Twitter Trends 2021. *Twitter Business Blog*, Dec. 2021. `https://marketing.x.com/pt/insights/the-conversation-twitter-trends-2021`. 2

[4] D. Brunet, E. R. Vrscay, and Z. Wang. On the mathematical properties of the structural similarity index. *IEEE Transactions on Image Processing*, 21(4):1488–1499, 2011. 15

[5] J. P. Cardenuto, J. Yang, R. Padilha, R. Wan, D. Moreira, H. Li, S. Wang, F. Andaló, S. Marcel, and A. Rocha. The age of synthetic realities: Challenges and opportunities. *APSIPA Transactions on Signal and Information Processing*, 12(1), 2023. 1

[6] V. De Silva and T. Sumanathilaka. A Survey on Image Captioning Using Object Detection and NLP. In *International Conference on Advanced Research in Computing (ICARC)*, pages 270–275. IEEE, 2024. 10

[7] P. Dhariwal and A. Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems (NIPS)*, 34:8780–8794, 2021. 8

[8] F. Dias, A. Moroni, and H. Pedrini. Using generative models to create a visual description of climate change. In *International Conference on Arts and Technology (ArtsIT), Interactivity and Game Creation*, pages 202–212. Springer, 2023. 2, 12

[9] H. Fang, S. Gupta, F. Iandola, R. K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J. C. Platt, C. L. Zitnick, and G. Zweig. From captions to visual concepts and back. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1473–1482, 2015. 12

[10] B. J. Fung, E. Sutlief, and M. G. H. Shuler. Dopamine and the interdependency of time perception and reward. *Neuroscience & Biobehavioral Reviews*, 125:380–391, 2021. 1

[11] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 10

[12] X. He and L. Deng. Deep learning for image-to-text generation: A technical overview. *IEEE Signal Processing Magazine*, 34(6):109–116, 2017. 11, 12

[13] P. D. Howe, M. Mildenberger, J. R. Marlon, and A. Leiserowitz. Geographic variation in opinions on climate change at state and local scales in the USA. *Nature Climate Change*, 5(6):596–603, 2015. 1

[14] D. P. Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 6

[15] L. F. Klein and M. K. Gold. *Debates in the Digital Humanities 2016*. University of Minnesota Press, 2016. 1

[16] R. Kraut, M. Patterson, V. Lundmark, S. Kiesler, T. Mukophadhyay, and W. Scherlis. Internet paradox: A social technology that reduces social involvement and psychological well-being? *American Psychologist*, 53(9):1017, 1998. 1

[17] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen. Deep learning for generic object detection: A survey. *International Journal of Computer Vision*, 128:261–318, 2020. 10

[18] V. Liu and L. B. Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Human Factors in Computing Systems (CHI)*, pages 1–23, 2022. 12

[19] Llama Team, AI @ Meta. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 7

[20] C. Lu and Y. Song. Simplifying, Stabilizing and Scaling Continuous-Time Consistency Models. *arXiv preprint arXiv:2410.11081*, 2024. 8

[21] Matt Novak. Why AI video won't work in Hollywood, according to a former Pixar animator. `https://qz.com/openai-sora-ai-video-hollywood-pixar-animation-1851442421`, 2024. Accessed: 2024-10-29. 5

[22] R. Mechrez, I. Talmi, and L. Zelnik-Manor. The contextual loss for image transformation with non-aligned data. In *European Conference on Computer Vision (ECCV)*, pages 768–783, 2018. 15

[23] A. Nichol, H. Jun, P. Dhariwal, P. Mishkin, and M. Chen. Point-E: A system for generating 3D point clouds from complex prompts. *arXiv preprint arXiv:2212.08751*, 2022. 8

[24] S. Niklaus, L. Mai, and F. Liu. Video frame interpolation via adaptive convolution. In *IEEE Conference on Computer Vision and Pattern Recognition (CPRV)*, pages 670–679, 2017. 16

[25] OpenAI. Video Generation Models as World Simulators. `https://openai.com/index/video-generation-models-as-world-simulators/`, 2024. Accessed: 2024-10-24. 6, 8

[26] J. Oppenlaender. A Taxonomy of Prompt Modifiers for Text-to-Image Generation. *arXiv preprint arXiv:2204.13988*, 2022. 12

[27] W. Peebles and S. Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023. 8

[28] A. Polyak, A. Zohar, A. Brown, A. Tjandra, A. Sinha, A. Lee, A. Vyas, B. Shi, C.-Y. Ma, C.-Y. Chuang, D. Yan, D. Choudhary, D. Wang, G. Sethi, G. Pang, H. Ma, I. Misra, J. Hou, J. Wang, K. Jagadeesh, K. Li, L. Zhang, M. Singh, M. Williamson, M. Le, M. Yu, M. K. Singh, P. Zhang, P. Vajda, Q. Duval, R. Girdhar, R. Sumbaly, S. S. Rambhatla, S. Tsai, S. Azadi, S. Datta, S. Chen, S. Bell, S. Ramaswamy, S. Sheynin, S. Bhattacharya, S. Motwani,

T. Xu, T. Li, T. Hou, W.-N. Hsu, X. Yin, X. Dai, Y. Taigman, Y. Luo, Y.-C. Liu, Y.-C. Wu, Y. Zhao, Y. Kirstain, Z. He, Z. He, A. Pumarola, A. Thabet, A. Sanakoyeu, A. Mallya, B. Guo, B. Araya, B. Kerr, C. Wood, C. Liu, C. Peng, D. Vengertsev, E. Schonfeld, E. Blanchard, F. Juefei-Xu, F. Nord, J. Liang, J. Hoffman, J. Kohler, K. Fire, K. Sivakumar, L. Chen, L. Yu, L. Gao, M. Georgopoulos, R. Moritz, S. K. Sampson, S. Li, S. Parmeggiani, S. Fine, T. Fowler, V. Petrovic, and Y. Du. Movie Gen: A Cast of Media Foundation Models. *arXiv preprint arXiv:2410.13720*, 2024. 6, 14

[29] Z. Qin, Q. Zeng, Y. Zong, and F. Xu. Image inpainting based on deep learning: A review. *Displays*, 69:102028, 2021. 14, 15

[30] J. Redmon. You only look once: Unified, real-time object detection. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 10

[31] M. C. Rehm and D. Jovanovic. Assembled Worlds: New Campo Marzio–Piranesi in the Age of AI. *Architectural Design*, 92(3):80–85, 2022. 6

[32] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(6):1137–1149, 2016. 10

[33] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 6

[34] A. Rougetet. Landscape Pictures Dataset. `https://www.kaggle.com/datasets/arnaud58/landscape-pictures`, 2020. Accessed: 2024-10-31. 15

[35] U. Saxena. Landscape Recognition Image Dataset - 12K Images. `https://www.kaggle.com/datasets/utkarshsaxenadn/landscape-recognition-image-dataset-12k-images`, 2021. Accessed: 2024-10-31. 15

[36] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages (VL)*, pages 336–343. IEEE, 1996. 2

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 10

[38] I. Skorokhodov, G. Sotnikov, and M. Elhoseiny. Aligning latent and image spaces to connect the unconnectable. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14144–14153, 2021. 15

[39] P. Sohn. Elemental. Pixar Animation Studios, 2023. 5

[40] Y. Song, P. Dhariwal, M. Chen, and I. Sutskever. Consistency models. *arXiv preprint arXiv:2303.01469*, 2023. 8

[41] S. Su, M. Delbracio, J. Wang, G. Sapiro, W. Heidrich, and O. Wang. Deep video deblurring for hand-held cameras. In *IEEE Conference on Computer Vision and Pattern Recognition (CPRV)*, pages 1279–1288, 2017. 16

[42] Y. Tay, M. Dehghani, V. Q. Tran, X. Garcia, J. Wei, X. Wang, H. W. Chung, S. Shakeri, D. Bahri, T. Schuster, H. S. Zheng, D. Zhou, N. Houlsby, and D. Metzler. Ul2: Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*, 2022. 7

[43] R. Thalheim. The Billion Dollar Code. Netflix, 2021. 2

[44] A. Tripathi. Impact of internet addiction on mental health: An integrative therapy is needed. *Integrative Medicine International*, 4(3-4):215–222, 2019. 1

[45] A. Villanueva, M. Ravella, B. Montell, T. Zhang, and H. Chang. Procedural Techniques for Large, Dynamic Sets in Elemental. In *ACM SIGGRAPH 2023 Talks*, pages 1–2, 2023. 5

[46] H. Xiang, Q. Zou, M. A. Nawaz, X. Huang, F. Zhang, and H. Yu. Deep learning for image inpainting: A survey. *Pattern Recognition*, 134:109046, 2023. 14, 15

[47] Xiaobai1217. Awesome Video Datasets. `https://github.com/xiaobai1217/Awesome-Video-Datasets`, 2023. Accessed: 2024-10-31. 15

[48] H. Xu, S. Xie, X. E. Tan, P.-Y. Huang, R. Howes, V. Sharma, S.-W. Li, G. Ghosh, L. Zettlemoyer, and C. Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 7

[49] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306, 2022. 7

[50] R. A. Yeh, C. Chen, T. Yian Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do. Semantic image inpainting with deep generative models. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5485–5493, 2017. 16

[51] J. Yu, Z. Lin, J. Yang, X. Shen, X. Lu, and T. S. Huang. Generative image inpainting with contextual attention. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5505–5514, 2018. 14

[52] Y. Zhang, T. T. Tzun, L. W. Hern, T. Sim, and K. Kawaguchi. Enhancing Semantic Fidelity in Text-to-Image Synthesis: Attention Regulation in Diffusion Models. *arXiv preprint arXiv:2403.06381*, 2024. 16

[53] Z. Zou, K. Chen, Z. Shi, Y. Guo, and J. Ye. Object detection in 20 years: A survey. *Proceedings of the IEEE*, 111(3):257–276, 2023. 10