ArtWalks via Latent Diffusion Models

ArtWalks via Latent Diffusion Models

Alberto Pennino
alberto.pennino@inf.ethz.ch
ETH Zürich
Zürich, Switzerland

Majed El Helou
majed.elhelou@inf.ethz.ch
ETH Zürich
Zürich, Switzerland

Daniel Vera Nieto
dveranieto@ethz.ch
ETH Zürich
Zürich, Switzerland

Fabio Zünd
fabio.zund@inf.ethz.ch
ETH Zürich
Zürich, Switzerland

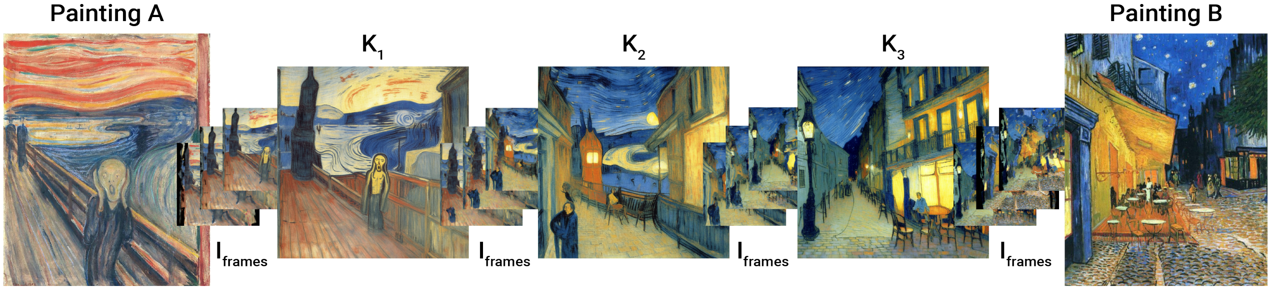

Figure 1: A complete video interpolation, or *walk*, between two paintings.

ABSTRACT

ArtWalks presents a creative application of multiple techniques in image and video generation. We explore the creation of a dynamic bridge between between two paintings, seamlessly transitioning between artworks adding motion to static images. We build a two-stage generative process, creating both abstract-conceptual interpolation as well as spatio-temporal interpolation. We first hallucinate intermediate artworks using a generative diffusion model, then interpolate between resulting frames using a motion generation network, thus obtaining a complete video out of static paintings.

CCS CONCEPTS

• **Applied computing → Fine arts**.

KEYWORDS

Generative Models, Motion Generation, Diffusion Models, Concept Interpolation, Image Interpolation , Applied Computer Vision, Art, Computational Art

ACM Reference Format:
Alberto Pennino, Majed El Helou, Daniel Vera Nieto, and Fabio Zünd. 2023. ArtWalks via Latent Diffusion Models. In *Proceedings of (ACM MIG '23).* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

ACM MIG '23, November 15–17, 2023, Rennes, France
© 2023 Association for Computing Machinery.
ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

Found within a museum, two paintings created by a masterful artist are displayed for observation. These images, frozen on canvas, spark a curiosity: what could unfold between these two moments? Could we explore a universe of unpainted artworks?

These questions drove our exploration and led to *ArtWalks*, a practical application of advanced techniques in image and video generation designed to amplify creative expression and to facilitate inventive modes of showcasing art. ArtWalks transforms static canvases into dynamic, visually engaging narratives through two approaches of interpolation. We first interpolate between abstract concepts that are present in paintings, then perform a spatio-temporal interpolation creating aesthetic and visually pleasing videos. With our approach each painting gradually evolves and merges into another. In what follows, we focus on the interpolation between two paintings. To extend this procedure to a sequence of multiple paintings, for instance, to an art collection, we repeat the A-to-B painting interpolation process in a chained fashion.

2 PROPOSED METHOD

Our proposed method, ArtWalks, consists of a two-stage generative process. In the conceptual interpolation stage, we employ a diffusion-based image generative model [Razzhigaev et al. 2023] to efficiently produce m intermediate images K_1 to K_m situated between two reference paintings, A and B. In the spatio-temporal interpolation stage, we employ a large motion interpolation network [Reda et al. 2022] to generate motion sequences between A, the generated images K and B. We finally upscale our results by employing a video super-resolution model [Wang et al. [n. d.]] reaching a resolution of 2K. The following section describes our method for transitioning, or *walking*, between two paintings. The pipeline is illustrated in Fig.2.

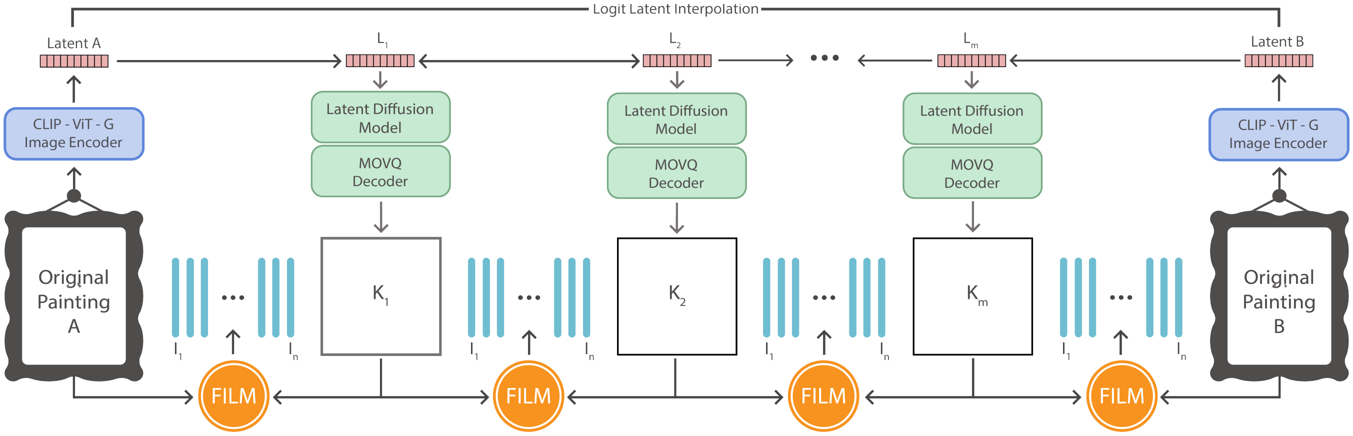

Figure 2: Overview of the ArtWalks pipeline, applied between two paintings.

2.1 Conceptual Interpolation

We begin by embedding both original paintings using a pre-trained CLIP-ViT-L14 image embedding model [Radford et al. 2021]. This process projects the paintings into an abstract latent space, preserving crucial information on their content and style.

To generate intermediate frames that conceptually transition from painting A to painting B, we interpolate between the two embeddings in the abstract latent space. This interpolation results in m intermediate embeddings L_1 to L_m, achieved through a linear combination of the two embeddings using a logit weighting function for controlling perceived speed. This choice ensures that the image generation remains consistent with a blend of both paintings.

2.2 Intermediate Image Generation

Having obtained all intermediate embeddings L_1 to L_m, we input them into an image generation model. In alignment with the approach proposed by [Razzhigaev et al. 2023], we employ a UNet model for latent diffusion generation [Rombach et al. 2021], followed by a MoVQGAN [Zheng et al. 2022] image decoder. Through the conditioning of the diffusion model with the interpolated image embeddings (Sec. 2.1), we steer the generative process towards producing images that encapsulate a conceptual and visual mixture of the original paintings.

2.3 Spatio-Temporal Interpolation

Following the previously outlined steps, we obtain intermediate images K_1 to K_m positioned between the original artworks A and B. When generating a video, it is crucial to ensure smooth transitions across all generated frames. To address this, we employ a FILM [Reda et al. 2022] frame interpolation network. Initially trained for interpolating between images with significant motion disparities, we adapt its application to our distinct context. FILM relies on a multi-scale feature extractor, which identifies shared content across frames, gradually adjusting objects and pixels across n interpolation frames I. This process results in a nuanced representation of motion for image elements between any two frames. By applying this procedure to all generated frames K_1 to K_m, we achieve a seamless video transition between paintings A and B.

2.4 Spatial and Temporal Resolution

In our approach, we place great emphasis on ensuring the final video exhibits both high spatial and temporal resolution. In contrast with a linear interpolation, we incorporate an S-curve approximation within the FILM interpolation process, ensuring natural, non-linear motion, therefore smoothing the transition between the K frames. This translates into a more immersive experience for viewers, as they are presented with a smoother, more dynamic, progression of visual elements.

Additionally, to achieve an optimal spatial resolution of 2K, we employ a state-of-the-art video super-resolution model [Wang et al. [n. d.]]. This model operates on the entire video, enhancing the overall visual quality by up-scaling details and improving the clarity of finer visual features such as brush strokes, highlights and reflections. This enhancement contributes significantly to the overall visual fidelity and ensures a more engaging visual experience.

3 EXPERIMENTAL RESULTS

In this section, we offer a visual comparison of the aforementioned interpolation techniques, each employed individually (see Fig.3). When solely relying on diffusion for interpolation, the resulting videos lack both motion and temporal coherence [Wang and Golland 2023]. On the other hand, applying the FILM model directly to paintings A and B yields smooth videos that suffer from spatial inconsistency. In contrast, ArtWalks leverages the strengths of both approaches, preserving both temporal and spatial consistency in the generated videos.

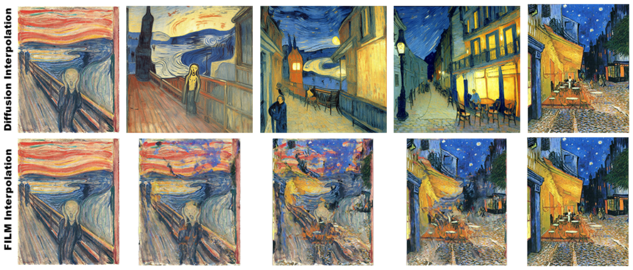

Figure 3: Application of diffusion and FILM interpolation.

4 CONCLUSION

We presented ArtWalks, a novel method to generate an artistic video out of a collection of paintings. Our interpolation takes place in the abstract conceptual domain as well as in the spatio-temporal domain. We improve the visual experience by carefully controlling the perceived speed, and we achieve a high spatial resolution. We hope that our work will help creators find further inspiration and allow viewers to experience art collections in a more dynamic and innovative way.

REFERENCES

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. *CoRR* abs/2103.00020 (2021). arXiv:2103.00020 https://arxiv.org/abs/2103.00020

Anton Razzhigaev, Arseniy Shakhmatov, Anastasia Maltseva, Vladimir Arkhipkin, Igor Pavlov, Ilya Ryabov, Angelina Kuts, Alexander Panchenko, Andrey Kuznetsov, and Denis Dimitrov. 2023. Kandinsky: an Improved Text-to-Image Synthesis with Image Prior and Latent Diffusion. arXiv:2310.03502 [cs.CV]

Fitsum Reda, Janne Kontkanen, Eric Tabellion, Deqing Sun, Caroline Pantofaru, and Brian Curless. 2022. FILM: Frame Interpolation for Large Motion. In *European Conference on Computer Vision (ECCV)*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *CoRR* abs/2112.10752 (2021). arXiv:2112.10752 https://arxiv.org/abs/2112.10752

Clinton J. Wang and Polina Golland. 2023. Interpolating between Images with Diffusion Models. arXiv:2307.12560 [cs.LG]

Xintao Wang, Liangbin Xie, Chao Dong, and Ying Shan. [n. d.]. Real-ESRGAN: Training Real-World Blind Super-Resolution with Pure Synthetic Data. In *International Conference on Computer Vision Workshops (ICCVW)* (2021).

Chuanxia Zheng, Long Tung Vuong, Jianfei Cai, and Dinh Phung. 2022. MoVQ: Modulating Quantized Vectors for High-Fidelity Image Generation. arXiv:2209.09002 [cs.CV]