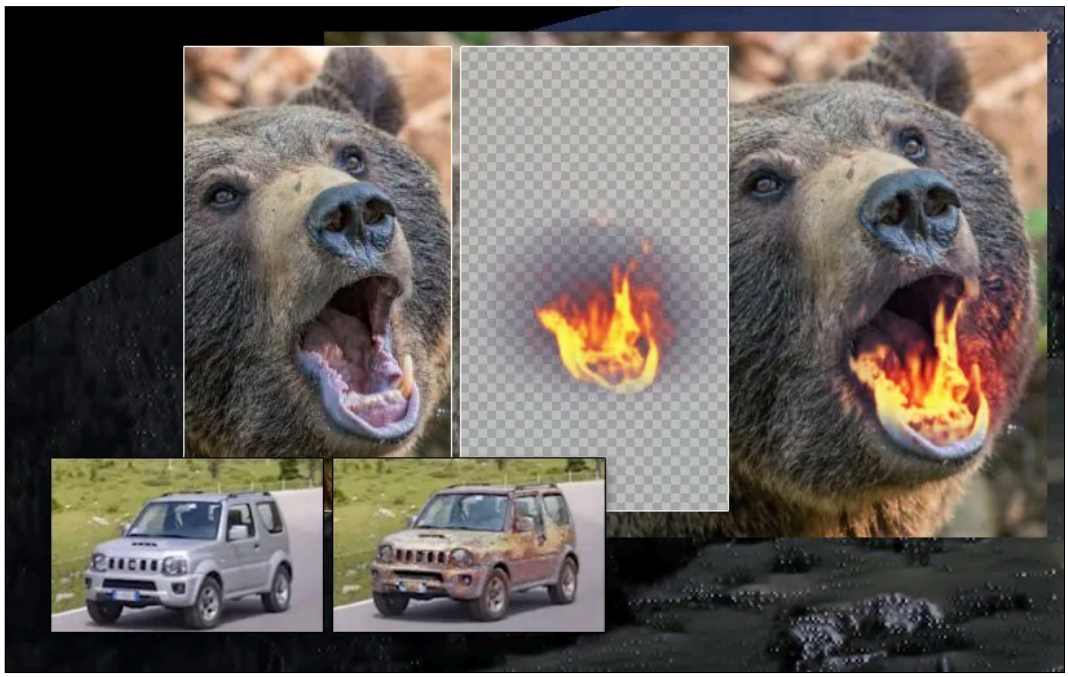
Consistent AI Video Content Editing with Text-Guided Input

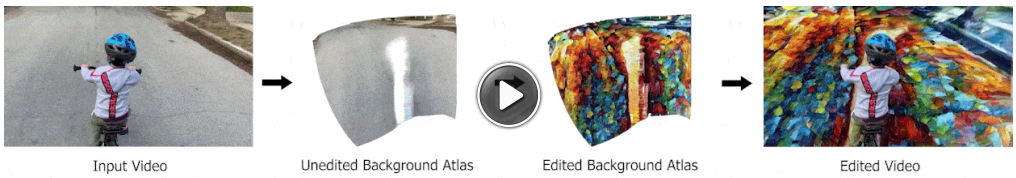
Originally published at https://www.unite.ai/consistent-ai-video-content-editing-with-text-guided-input/



While the professional VFX community is intrigued – and occasionally feels a little threatened – by new innovations in image and video synthesis, the lack of temporal continuity in most AI-based video editing projects relegates many of these efforts to the 'psychedelic' sphere, with shimmering and rapidly altering textures and structures, inconsistent effects and the kind of crude technology-wrangling that recalls the photochemical era of visual effects.

If you want to change something very specific in a video that doesn't fall into the realm of deepfakes (i.e., imposing a new identity on existing footage of a person), most of the current solutions operate under quite severe limitations, in terms of the precision required for production-quality visual effects.

One exception is the ongoing work of a loose association of academics from the Weizmann Institute of Science. In 2021, three of its researchers, in association with Adobe , announced a novel method for decomposing video and superimposing a consistent internal mapping – a *layered neural atlas* – into a composited output, complete with alpha channels and temporally cohesive output.



CLICK TO VIEW ANIMATED GIF

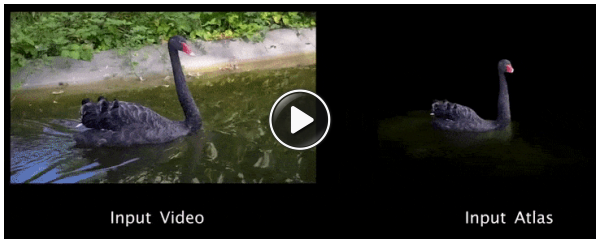
From the 2021 paper: an estimation of the complete traversal of the road in the source clip is edited via a neural network in a manner which traditionally would req moving. Since the background and foreground elements are handled by different networks, masks are truly 'automatic'. Source: https://layered-neural-atlases.gith

Though it falls somewhere into the realm covered by optical flow in VFX pipelines, the layered atlas has no direct equivalent in traditional CGI workflows, since it essentially constitutes a 'temporal texture map' that can be produced and edited through traditional software methods. In the second image in the illustration above, the background of the road surface is represented (figuratively) across the entire runtime of the video. Altering that base image (third image from left in the illustration above) produces a consistent change in the background.

The images of the 'unfolded' atlas above only represent individual interpreted frames; consistent changes in any target video frame are mapped back to the original frame, retaining any necessary occlusions and other requisite scene effects, such as shadows or reflections.

The core architecture uses a Multilayer Perceptron (MLP) to represent the unfolded atlases, alpha channels and mappings, all of which are optimized in concert, and entirely in a 2D space, obviating NeRF-style prior knowledge of 3D geometry points, depth maps, and similar CGI-style trappings.

The reference atlas of individual objects can also be reliably altered:



CLICK TO VIEW ANIMATED GIF

Consistent change to a moving object under the 2021 framework. Source: https://www.youtube.com/watch?v=aQhakPFC4oQ

Essentially the 2021 system combines geometry alignment, match-moving, mapping, re-texturizing and rotoscoping into a discrete neural process.

Text2Live

The three original researchers of the 2021 paper, together with NVIDIA research, are among the contributors to a new innovation on the technique that combines the power of layered atlases with the kind of text-guided CLIP technology that has come back to prominence this week with OpenAI's release of the DALL-E 2 framework.

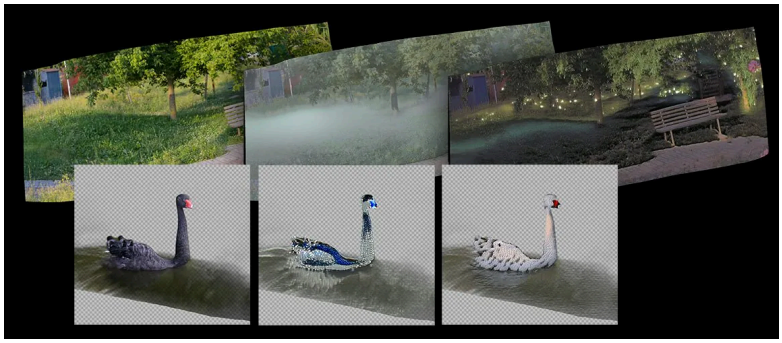
The new architecture, titled *Text2Live*, allows an end user to create localized edits to actual video content based on text prompts:

CLICK TO VIEW ANIMATED GIF

CLICK TO VIEW ANIMATED GIF

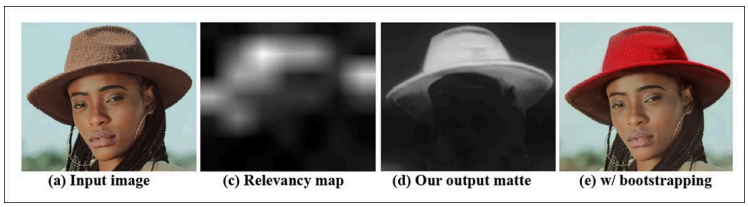
Two examples of foreground editing. For better resolution and definition, check out the original videos at https://text2live.github.io/sm/pages/video_results_atlases.html

Text2Live offers semantic and highly localized editing without the use of a pre-trained generator, by making use of an internal database that's specific to the video clip being affected.



Background and foreground (object) transformations under Text2Live. Source: https://text2live.github.io/sm/pages/video_results_atlases.html

The technique does not require user-provided masks, such as a typical rotoscoping or green-screen workflow, but rather estimates *relevancy maps* through a bootstrapping technique based on 2021 research from The School of Computer Science at Tel Aviv University and Facebook AI Research (FAIR).

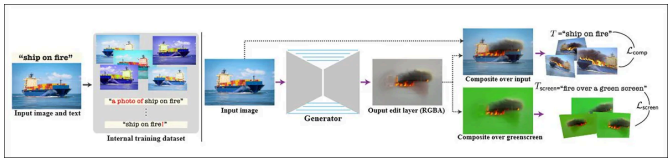


Output maps generated via a transformer-based generic attention model.

The new paper is titled *Text2LIVE: Text-Driven Layered Image and Video Editing*. The original 2021 team is joined by Weizmann's Omer Bar-Tal, and Yoni Kasten of NVIDIA Research.

Architecture

Text2Live comprises a generator trained on a sole input image and target text prompts. A Contrastive Language-Image Pretraining (CLIP) model pre-trained on 400 million text/image pairs provides associated visual material from which user-input transformations can be interpreted.



The generator accepts an input image (frame) and outputs a target RGBA layer containing color and opacity information. This layer is then composited into the original footage with additional augmentations.

The alpha channel in the generated RGBA layer provides an internal compositing function without recourse to traditional pipelines involving pixel-based software such as After Effects.

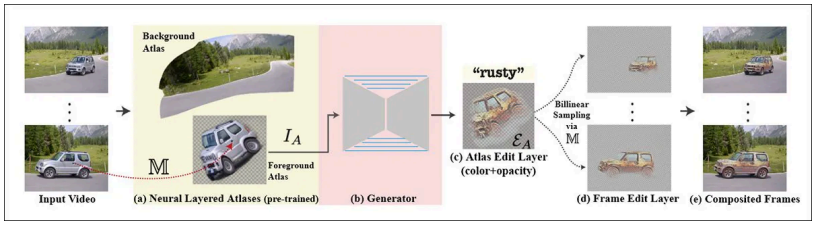
By training on internal images relevant to the target video or image, Text2Live avoids the requirement either to invert the input image into the latent space of a Generative Adversarial Network (GAN), a practice which is currently far from exact enough for production video editing requirements, or else use a Diffusion model that's more precise and configurable, but cannot maintain fidelity to the target video.

Sundry prompt-based transformation edits from Text2Live.

Prior approaches have either used propagation-based methods or optical flow-based approaches. Since these techniques are to some or other extent frame-based, neither is capable of creating a consistent temporal appearance of changes in output video. A neural layered atlas, instead, provides a single space in which to address changes, which can then remain faithful to the committed change as the video progresses.

CLICK TO VIEW ANIMATED GIF

No 'sizzling' or random hallucinations: Text2Live obtains an interpretation of the text prompt 'rusty jeep', and applies it once to the neural layered atlas of the car in the video, instead of restarting the transformation for each interpreted frame.



Workflow of Text2Live's consistent transformation of a Jeep into a rusty relic.

Text2Live is closer to a breakthrough in AI-based compositing, rather than in the fertile text-to-image space which has attracted so much attention this week with the release of the second generation of OpenAI's DALL-E framework (which can incorporate target images as part of the transformative process, but remains limited in its ability to directly intervene in a photo, in addition to the censoring of source training data and imposition of filters, designed to prevent user abuse).

Rather, Text2Live allows the end user to extract an atlas and then edit it in one pass in high-control pixel-based environments such as Photoshop (and arguably even more abstract image synthesis frameworks such as NeRF), before feeding it back into a correctly-oriented environment that nonetheless does not rely on 3D estimation or backwards-looking CGI-based approaches.

Furthermore, Text2Live, the authors claim, is the first comparable framework to achieve masking and compositing in an entirely automatic manner.

First published 7th April 2022.

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More