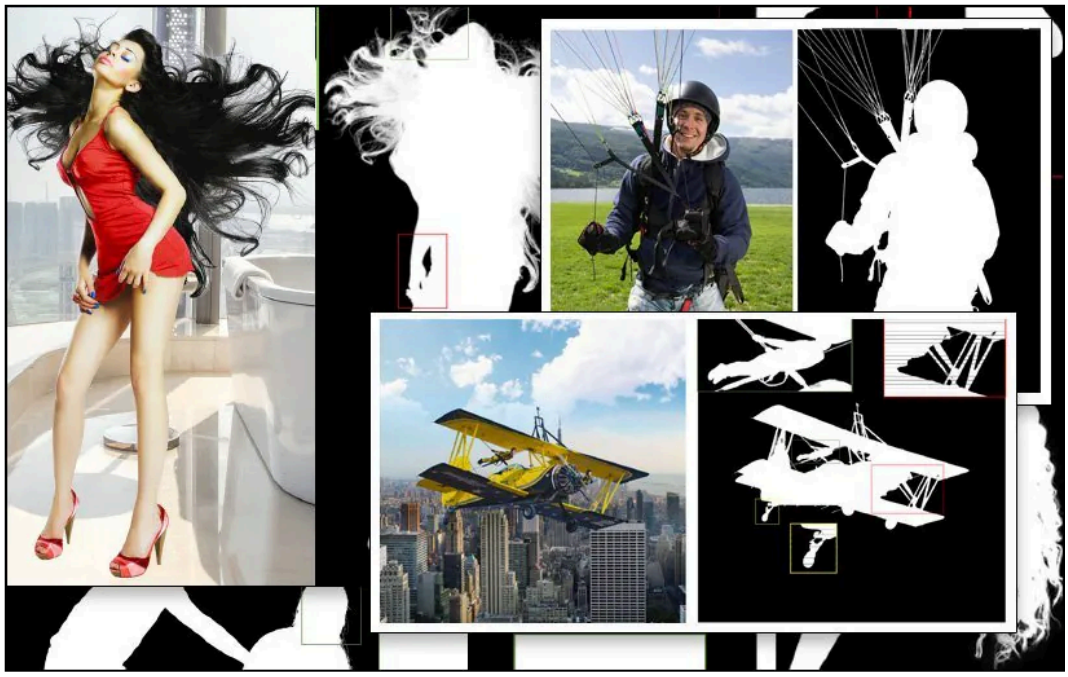
AI Image Matting That Understands Scenes

Originally published at https://www.unite.ai/ai-image-matting-that-understands-scenes/



In the extras documentary accompanying the 2003 DVD release of *Alien*^{3} (1992), visual effects legend Richard Edlund recalled with horror the 'sumo wrestling' of photochemical matte extraction that dominated visual effects work between the late 1930s and the late 1980s. Edlund described the hit-and-miss nature of the process as 'sumo wrestling', in comparison to the digital blue/green-screen techniques that took over in the early 1990s (and he has returned to the metaphor since).

Extracting a foreground element (such as a person or a spaceship model) from a background, so that the cut-out image can be composited into a background plate, was originally achieved by filming the foreground object against a uniform blue or green background.

CLICK TO VIEW ANIMATED GIF

Laborious photochemical extraction processes for a VFX shot by ILM for 'Return of the Jedi' (1983). Source: https://www.youtube.com/watch?v=qwMLOjqPmbQ

In the resulting footage, the background color would subsequently be isolated chemically and used as a template to reprint the foreground object (or person) in an optical printer as a 'floating' object in an otherwise transparent film cell.

The process was known as color separation overlay (CSO) – though this term would eventually become more associated with the crude 'Chromakey' video effects in lower-budgeted television output of the 1970s and 1980s, which were achieved with analogue rather than chemical or digital means.

A demonstration of Color Separation Overlay in 1970 for the British children's show 'Blue Peter'. Source: https://www.bbc.co.uk/archive/blue_peter_noakes_CSO/zwb9vwx

In any case, whether for film or video elements, thereafter the extracted footage could be inserted into any other footage.

Though Disney's notably more expensive and proprietary sodium-vapor process (which keyed on yellow, specifically, and was also used for Alfred Hitchcock's 1963 horror *The Birds*) gave better definition and crisper mattes, photochemical extraction remained painstaking and unreliable.

Disney's proprietary sodium vapor extraction process required backgrounds near the yellow end of the spectrum. Here, Angela Lansbury is suspended on wires during the production of a VFX-laced sequence for 'Bedknobs and Broomsticks' (1971). Source

Beyond Digital Matting

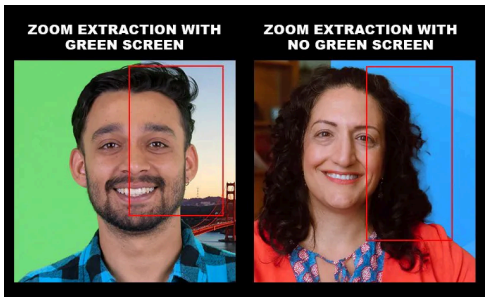
In the 1990s, the digital revolution dispensed with the chemicals, but not the need for green screens. It was now possible to remove the green (or whatever color) background just by searching for pixels within a tolerance range of that color, in pixel-editing software such as Photoshop, and a new generation of video-compositing suites that could automatically key out the colored backgrounds. Almost overnight, sixty years of the optical printing industry were consigned to history.

The last ten years of GPU-accelerated computer vision research is ushering matte extraction into a third age, tasking researchers with the development of systems that can extract high-quality mattes without the need for green screens. At Arxiv alone, papers related to innovations in machine learning-based foreground extraction are a weekly feature.

Putting Us in the Picture

This locus of academic and industry interest in AI extraction has already impacted the consumer space: crude but workable implementations are familiar to us all in the form of Zoom and Skype filters that can replace our living-room backgrounds with tropical islands, et al, in video conference calls.

However, the best mattes still require a green screen, as Zoom noted last Wednesday.



Left, a man in front of a green screen, with well-extracted hair via Zoom's Virtual Background feature. Right, a woman in front of a normal domestic scene, with hair extracted algorithmically, less accurately, and with higher computing requirements. Source: https://support.zoom.us/hc/en-us/articles/210707503-Changing-your-Virtual-Background-image

A further post from the Zoom Support platform warns that non-green-screen extraction also requires greater computing power in the capture device.

The Need to Cut It Out

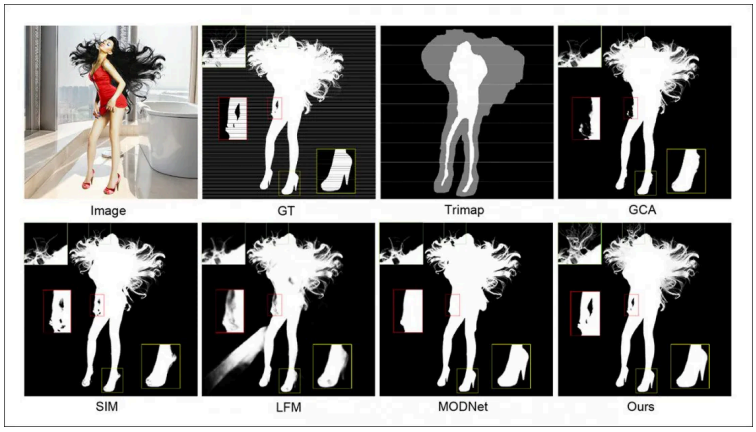
Improvements in quality, portability and resource economy for 'in the wild' matte extraction systems (i.e. isolating people without the need for green screens) are relevant to many more sectors and pursuits than just videoconferencing filters.

For dataset development, improved facial, full-head and full-body recognition offers the possibility of ensuring that extraneous background elements don't get trained into computer vision models of human subjects; more accurate isolation would greatly improve semantic segmentation techniques designed to distinguish and assimilate domains (i.e. *'cat'*, *'person'*, *'boat'*), and improve VAE and transformer-based based image synthesis systems such as OpenAI's new DALL-E 2; and better extraction algorithms would cut down on the need for expensive manual rotoscoping in costly VFX pipelines.

In fact, the ascendancy of multimodal (usually text/image) methodologies, where a domain such as 'cat' is encoded both as an image and with associated text references, is already making inroads into image processing. One recent example is the Text2Live architecture, which uses multimodal (text/image) training to create videos of, among myriad other possibilities, crystal swans and glass giraffes.

Scene-Aware AI Matting

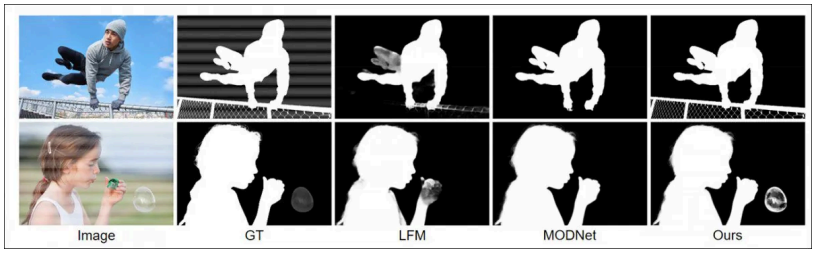
A good deal of research into AI-based automatic matting has focused on boundary recognition and evaluation of pixel-based groupings inside an image or video frame. However, new research from China offers an extraction pipeline that improves delineation and matte quality by leveraging *text-based descriptions* of a scene (a multimodal approach that has gained traction in the computer vision research sector over the last 3-4 years), claiming to have improved on prior methods in a number of ways.



An example SPG-IM extraction (last image, lower right), compared against competing prior methods. Source: https://arxiv.org/pdf/2204.09276.pdf

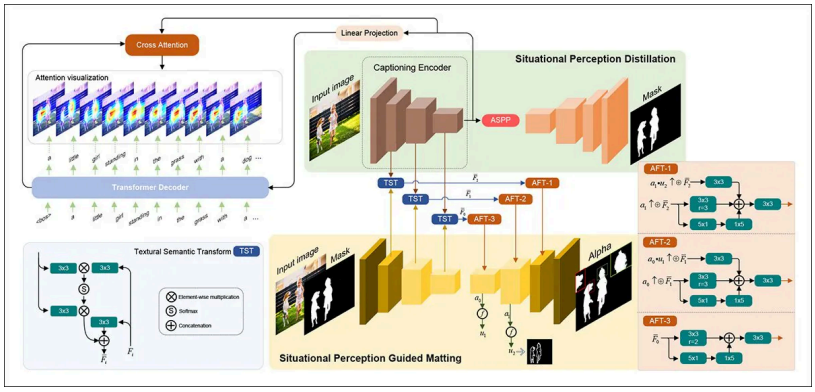
The challenge posed for the extraction research sub-sector is to produce workflows that require a bare minimum of manual annotation and human intervention – ideally, none. Besides the cost implications, the researchers of the new paper observe that annotations and manual segmentations undertaken by outsourced crowdworkers across various cultures can cause images to be labeled or even segmented in different ways, leading to inconsistent and unsatisfactory algorithms.

One example of this is the subjective interpretation of what defines a 'foreground object':



From the new paper: prior methods LFM and MODNet ('GT' signifies Ground Truth, an 'ideal' result often achieved manually or by non-algorithmic methods), have takes on the definition of foreground content, whereas the new SPG-IM method more effectively delineates 'near content' through scene context.

To address this, the researchers have developed a two-stage pipeline titled *Situational Perception Guided Image Matting* (SPG-IM). The two-stage encoder/decoder architecture comprises Situational Perception Distillation (SPD) and Situational Perception Guided Matting (SPGM).



The SPG-IM architecture.

First, SPD pretrains visual-to-textual feature transformations, generating captions apposite to their associated images. After this, the foreground mask prediction is enabled by connecting the pipeline to a novel saliency prediction technique.

Then SPGM outputs an estimated alpha matte based on the raw RGB image input and the generated mask obtained in the first module.

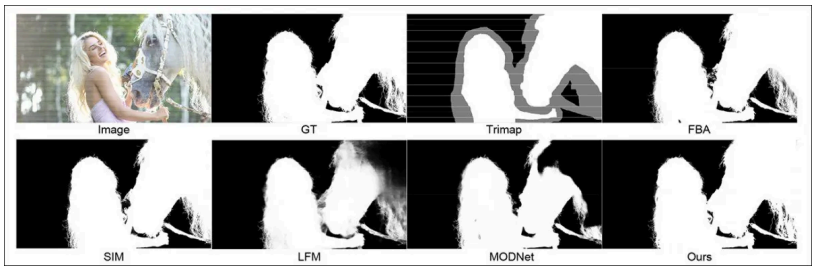
The objective is situational perception guidance, wherein the system has a contextual understanding of what the image consists of, allowing it to frame – for instance- the challenge of extracting complex hair from a background against known characteristics of such a specific task.

In the example below, SPG-IM understands that the cords are intrinsic to a 'parachute', where MODNet fails to retain and define these details. Likewise above, the complete structure of the playground apparatus is arbitrarily lost in MODNet.

The new paper is titled *Situational Perception Guided Image Matting*, and comes from researchers at the OPPO Research Institute, PicUp.ai, and Xmotors.

Intelligent Automated Mattes

SPG-IM also proffers an Adaptive Focal Transformation (AFT) Refinement Network that can process local details and global context separately, facilitating 'intelligent mattes'.

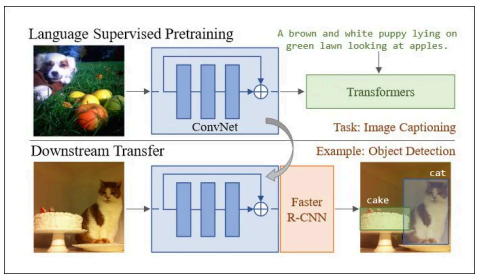


Understanding scene context, in this case 'girl with horse', can potentially make foreground extraction easier than prior methods.

The paper states:

> *'We believe that visual representations from the visual-to-textual task, e.g. image captioning, focus on more semantically comprehensive signals between a)object to object and b)object to the ambient environment to generate descriptions that can cover both the global info and local details. In addition, compared with the expensive pixel annotation of image matting, textual labels can be massively collected at a very low cost.'*

The SPD branch of the architecture is jointly pretrained with the University of Michigan's VirTex transformer-based textual decoder, which learns visual representations from semantically dense captions.



VirTex jointly trains a ConvNet and Transformers via image-caption couplets, and transfers the obtained insights to downstream vision tasks such as object detection. Source: https://arxiv.org/pdf/2006.06666.pdf

Among other tests and ablation studies, the researchers tested SPG-IM against state-of-the-art trimap-based methods Deep Image Matting (DIM), IndexNet, Context-Aware Image Matting (CAM), Guided Contextual Attention (GCA), FBA, and Semantic Image Mapping (SIM).

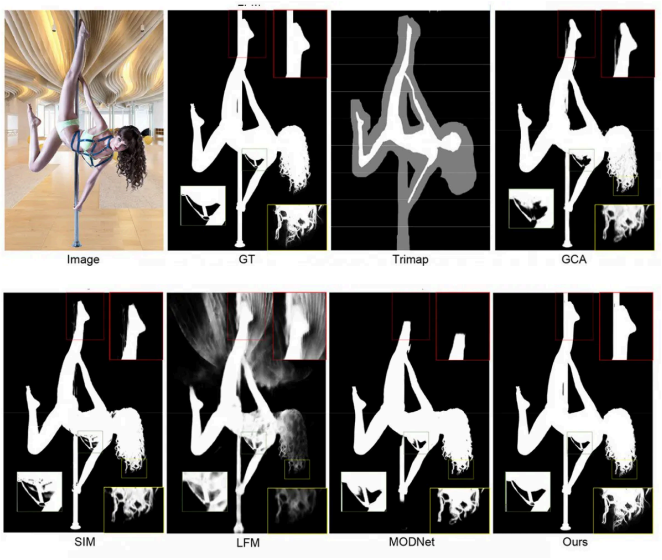
Other prior frameworks tested included trimap-free approaches LFM, HAttMatting, and MODNet. For fair comparison, the test methods were adapted based on the differing methodologies; where code was not available, the paper's techniques were reproduced from the described architecture.

The new paper states:

> *'Our SPG-IM outperforms all competing trimap-free methods ([LFM], [HAttMatting], and [MODNet]) by a large margin. Meanwhile, our model also shows remarkable superiority over the state-of-the-art (SOTA) trimap-based and mask-guided methods in terms of all four metrics across the public datasets (i.e. Composition-1K, Distinction-646, and Human-2K), and our Multi-Object-1K benchmark.'*

And continues:

'It can be obviously observed that our method preserves fine details (e.g. hair tip sites, transparent textures, and boundaries) without the guidance of trimap. Moreover, compared to other competing trimap-free models, our SPG-IM can retain better global semantic completeness.'



First published 24th 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