Removing Objects From Video More Efficiently With Machine Learning

Originally published at https://www.unite.ai/removing-objects-from-video-more-efficiently-with-machine-learning/

New research from China reports state-of-the-art results – as well as an impressive improvement in efficiency – for a new video inpainting system that can adroitly remove objects from footage.

CLICK TO VIEW ANIMATED GIF

A hang-glider's harness is painted out by the new procedure. See the source video for better resolution and more examples.
Source: https://www.youtube.com/watch?v=N–qC3T2wc4

The technique, called End-to-End framework for Flow-Guided video Inpainting (*E^{2}FGVI*), is also capable of removing watermarks and various other kinds of occlusion from video content.

CLICK TO VIEW ANIMATED GIF

E2FGVI calculates predictions for content that lies behind occlusions, enabling the removal of even notable and otherwise intractable watermarks. Source: https://github.com/MCG-NKU/E2FGVI

(To see more examples in better resolution, check out the video)

Though the model featured in the published paper was trained on 432px x 240px videos (commonly low input sizes, constrained by available GPU space vs. optimal batch sizes and other factors), the authors have since released *E^{2}FGVI-HQ*, which can handle videos at an arbitrary resolution.

The code for the current version is available at GitHub, while the HQ version, released last Sunday, can be downloaded from Google Drive and Baidu Disk.

CLICK TO VIEW ANIMATED GIF

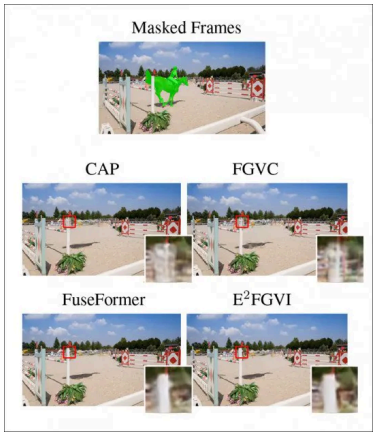
The kid stays in the picture.

E^{2}FGVI can process 432×240 video at 0.12 seconds per frame on a Titan XP GPU (12GB VRAM), and the authors report that the system operates fifteen times faster than prior state-of-the-art methods based on optical flow.

CLICK TO VIEW ANIMATED GIF

A tennis player makes an unexpected exit.

Tested on standard datasets for this sub-sector of image synthesis research, the new method was able to outperform rivals in both qualitative and quantitative evaluation rounds.



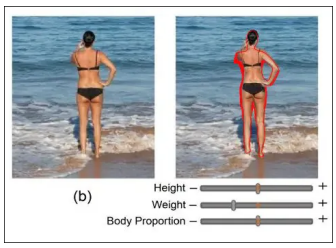
Tests against prior approaches. Source: https://arxiv.org/pdf/2204.02663.pdf

The paper is titled *Towards An End-to-End Framework for Flow-Guided Video Inpainting*, and is a collaboration between four researchers from Nankai University, together with a researcher from Hisilicon Technologies.

What's Missing in This Picture

Besides its obvious applications for visual effects, high quality video inpainting is set to become a core defining feature of new AI-based image synthesis and image-altering technologies.

This is particularly the case for body-altering fashion applications, and other frameworks that seek to 'slim down' or otherwise alter scenes in images and video. In such cases, it's necessary to convincingly 'fill in' the extra background that is exposed by the synthesis.



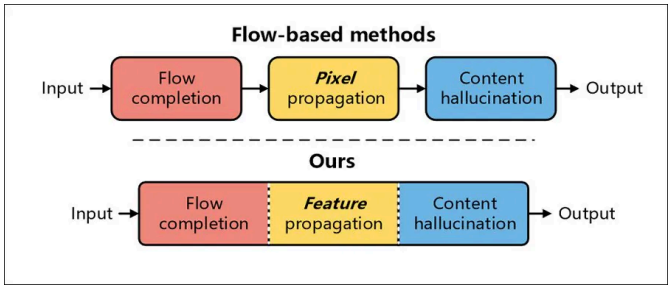
From a recent paper, a body 'reshaping' algorithm is tasked with inpainting the newly-revealed background when a subject is resized. Here, that shortfall is represented by the red outline that the (real life, see image left) fuller-figured person used to occupy. Based on source material from https://arxiv.org/pdf/2203.10496.pdf

Coherent Optical Flow

Optical flow (OF) has become a core technology in the development of video object removal. Like an atlas, OF provides a one-shot map of a temporal sequence. Often used to measure velocity in computer vision initiatives, OF can also enable temporally consistent in-painting, where the aggregate sum of the task can be considered in a single pass, instead of Disney-style 'per-frame' attention, which inevitably leads to temporal discontinuity.

Video inpainting methods to date have centered on a three-stage process: *flow completion*, where the video is essentially mapped out into a discrete and explorable entity; *pixel propagation*, where the holes in 'corrupted' videos are filled in by bidirectionally propagating pixels; and *content hallucination* (pixel 'invention' that's familiar to most of us from deepfakes and text-to-image frameworks such as the DALL-E series) where the estimated 'missing' content is invented and inserted into the footage.

The central innovation of E^{2}FGVI is to combine these three stages into an end-to-end system, obviating the need to carry out manual operations on the content or the process.

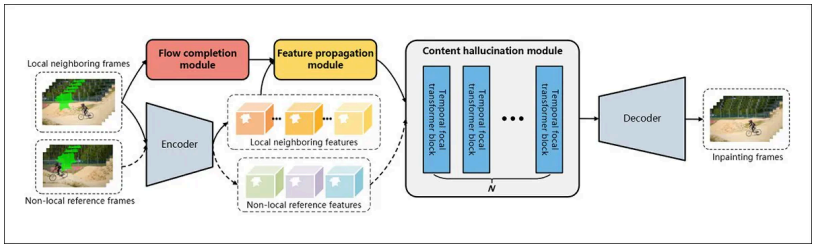


The paper observes that the need for manual intervention requires that older processes not take advantage of a GPU, making them quite time-consuming. From the paper*:

> *'Taking DFVI as an example, completing one video with the size of 432 × 240 from DAVIS, which contains about 70 frames, needs about 4 minutes, which is unacceptable in most real-world applications. Besides, except for the above-mentioned drawbacks, only using a pretrained image inpainting network at the content hallucination stage ignores the content relationships across temporal neighbors, leading to inconsistent generated content in videos.'*

By uniting the three stages of video inpainting, E^{2}FGVI is able to substitute the second stage, pixel propagation, with feature propagation. In the more segmented processes of prior works, features are not so extensively available, because each stage is relatively hermetic, and the workflow only semi-automated.

Additionally, the researchers have devised a *temporal focal transformer* for the content hallucination stage, which considers not just the direct neighbors of pixels in the current frame (i.e. what is happening in that part of the frame in the previous or next image), but also the distant neighbors that are many frames away, and yet will influence the cohesive effect of any operations performed on the video as a whole.



Architecture of E2FGVI.

The new feature-based central section of the workflow is able to take advantage of more feature-level processes and learnable sampling offsets, while the project's novel focal transformer, according to the authors, extends the size of focal windows 'from 2D to 3D'.

Tests and Data

To test E^{2}FGVI, the researchers evaluated the system against two popular video object segmentation datasets: YouTube-VOS, and DAVIS. YouTube-VOS features 3741 training video clips, 474 validation clips, and 508 test clips, while DAVIS features 60 training video clips, and 90 test clips.

E^{2}FGVI was trained on YouTube-VOS and evaluated on both datasets. During training, object masks (the green areas in the images above, and the accompanying YouTube video) were generated to simulate video completion.

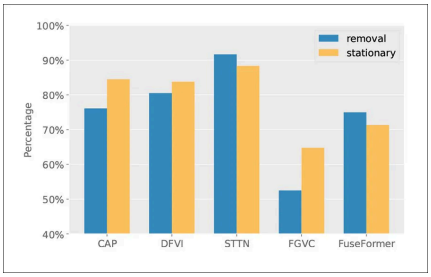
For metrics, the researchers adopted Peak signal-to-noise ratio (PSNR), Structural similarity (SSIM), Video-based Fréchet Inception Distance (VFID), and Flow Warping Error – the latter to measure temporal stability in the affected video.

The prior architectures against which the system was tested were VINet, DFVI, LGTSM, CAP, FGVC, STTN, and FuseFormer.

	Accuracy								Efficiency	
	YouTube-VOS				DAVIS				FLOPs	Runtime (s/frame)
Models	PSNR ↑	SSIM ↑	VFID ↓	E_{warp}^* ↓	PSNR ↑	SSIM ↑	VFID ↓	E_{warp}^* ↓		
VINet [23]	29.20	0.9434	0.072	0.1490	28.96	0.9411	0.199	0.1785	-	-
DFVI [57]	29.16	0.9429	0.066	0.1509	28.81	0.9404	0.187	0.1608	-	2.56
LGTSM [8]	29.74	0.9504	0.070	0.1859	28.57	0.9409	0.170	0.1640	1008G	0.23
CAP [28]	31.58	0.9607	0.071	0.1470	30.28	0.9521	0.182	0.1533	861G	0.40
FGVC [17]	29.67	0.9403	0.064	0.1022	30.80	0.9497	0.165	0.1586	-	2.44
STTN [63]	32.34	0.9655	0.053	0.0907	30.67	0.9560	0.149	0.1449	1032G	0.12
FuseFormer [33]	33.29	0.9681	0.053	0.0900	32.54	0.9700	0.138	0.1362	752G	0.20
E^{2}FGVI (Ours)	**33.71**	**0.9700**	**0.046**	**0.0864**	**33.01**	**0.9721**	**0.116**	**0.1315**	682G	0.16

From the quantitative results section of the paper. Up and down arrows indicate that higher or lower numbers are better, respectively. E2FGVI achieves the best s methods are evaluated according to FuseFormer, though DFVI, VINet and FGVC are not end-to-end systems, making it impossible to estimate their FLOPs.

In addition to achieving the best scores against all competing systems, the researchers conducted a qualitative user-study, in which videos transformed with five representative methods were shown individually to twenty volunteers, who were asked to rate them in terms of visual quality.



The vertical axis represents the percentage of participants that preferred the E^{2}FGVI output in terms of visual quality.

The authors note that in spite of the unanimous preference for their method, one of the results, FGVC, does not reflect the quantitative results, and they suggest that this indicates that E^{2}FGVI might, speciously, be generating 'more visually pleasant results'.

In terms of efficiency, the authors note that their system greatly reduces floating point operations per second (FLOPs) and inference time on a single Titan GPU on the DAVIS dataset, and observe that the results show E^{2}FGVI running x15 faster than flow-based methods.

They comment:

> *'[E^{2}FGVI] holds the lowest FLOPs in contrast to all other methods. This indicates that the proposed method is highly efficient for video inpainting.'*

**My conversion of authors' inline citations to hyperlinks.*

First published 19th May 2022.

Amended Tuesday 28th October 2025, to remove faulty video embed and amend references to embedded video in the article body.

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