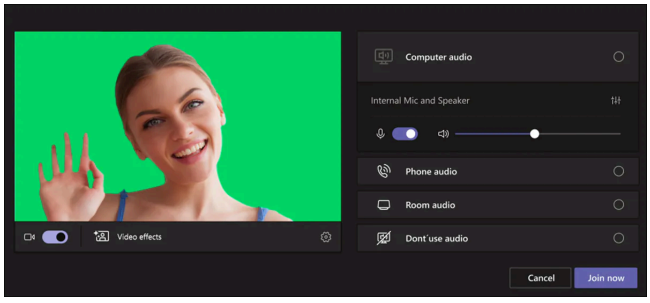
Improving AI Background Removal Without Costly Human Annotation

Originally published at https://www.unite.ai/improving-ai-background-removal-without-costly-human-annotation/

New research shows AI can cleanly cut people out of video without expensive human labeling, improving quality and stability

Most of us will have experienced being 'dropped out' of a background through simple background-altering/obscuring filters on video chat platforms – and we may have noticed the limitations of such systems, which are trained on the most frequent types of cases likely to be found in a video conference, and are not usually robust to anything 'unexpected' – or even to *predictable* objects, such as fingers:



A typical example of an AI-trained foreground extraction system cropping out too much of the source subject. Source

The easiest solution, and one which is becoming increasingly popular in the face of vision language models (VLMs) that are not specifically trained for such tasks, is to fine-tune an existing model on data likely to be encountered by the system:

Two high-volume, excruciatingly-annotated datasets underpin the solutions offered in the 2020 paper 'Real-Time High-Resolution Background Matting'. Source

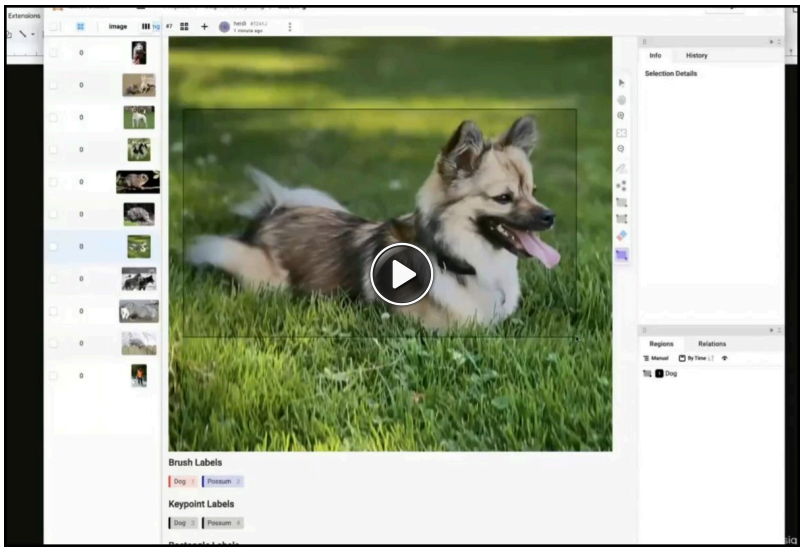
However, such data has to be annotated, at some expense and at some cost of time, by humans; and in any case, this results in a very specific and non-generalized tool that costs a lot of money, and can't usually be used for a wider variety of tasks.

Nonetheless, developing ‘targeted’ models of this type is currently the shortest route to effective inference in a variety of domains, not just video and image matting (i.e., background removal/foreground mattes). To date, many of the unsupervised solutions offered have more or less just pushed the problem around on the plate.

Even now, despite the proliferation of AI-augmented filters in platforms such as Zoom, the latter continues to recommend a green screen as the optimal solution for background removal – a burdensome and somewhat ‘professional’ solution that would perhaps make most of us a little self-conscious.

Cut It Out!

Lately, interest has grown in using the Segment Anything (SAM) family to provide automated and fine-grained extraction. Since SAM was developed to aid annotation rather than provide crisp outlines to a standard acceptable in a visual effects pipeline, its default boundaries are not suitable to address the challenge, unaided:



CLICK TO PLAY VIDEO ON SITE

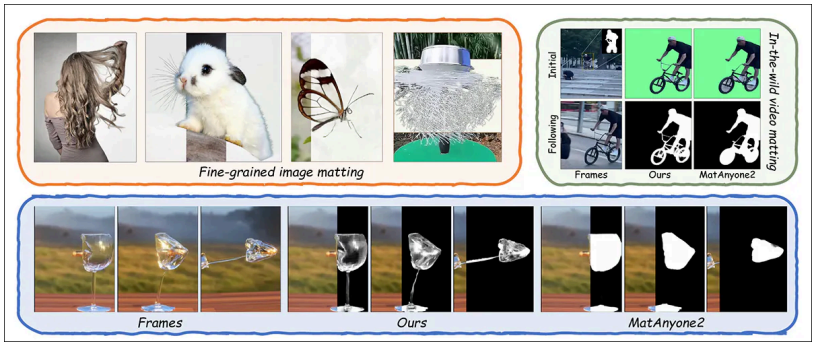
Click to play, as necessary. *An example of the rough borders created by a Segment Anything model – ideal for annotation, but not good enough for VFX drop-out.* Source

A recent offering from China has proposed a more sophisticated way of using SAM models to obtain superior extraction processes – by marrying a foundational tracker such as SAM to a region-proposal bridge with dedicated matting heads. In this way, the system is able to refine edge detail iteratively and resolve challenging edges, such as hair in motion:

CLICK TO PLAY VIDEO ON SITE

Click to play, as necessary. *From the ancillary site supporting the new paper, an amalgam of supplementary videos, demonstrating the sophistication of the authors’ method for extraction.* Source

Crucially, the new composite system trains only on images, not on videos, and requires no additional human annotation – the traditional hindrance against progress in this, and diverse other AI domains.



Examples of fine-grained image and video matting achieved with the new method, with challenging cases such as hair, transparency, and motion shown alongside the method produces – the authors contend – cleaner, more stable results than prior approaches. Source

With three experimental models produced for the work, the authors claim new state-of-the-art performance in this task, while retaining the higher generalization capabilities of the base model, meaning that the method produces an all-purpose model with extra capabilities, rather than a siloed tool targeted to a single task.

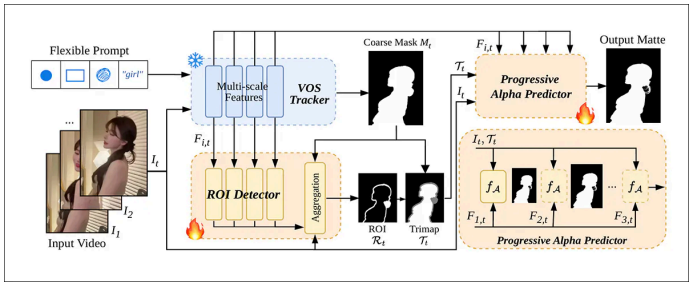
The authors state:

> *'Comprehensive experiments show that SAM2Matting achieves state-of-the-art (SOTA) performance on both image and video matting, with video matting evaluated in a strictly zero-shot manner.*
>
> *'Extensive in-the-wild results further demonstrate its strong generalization to open-world scenarios with rapid motion, complex backgrounds, and target attachments (e.g., man riding a bicycle).*
>
> *'Moreover, our matting components are lightweight and efficient, enabling the SAM2.1-Tiny variant to run at 40 FPS on a 200-frame 1080p video using less than 5GB GPU memory.'*

The new paper is titled *SAM2Matting: Generalized Image and Video Matting*, and comes from four authors across Fudan University and Shanghai University of Finance and Economics. The work has a GitHub repository, which at the time of writing has released checkpoints of different variants, inference code, and an interactive demo, with a release of training code promised. Additionally, there is a project site.

Method

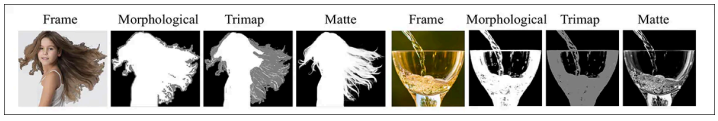
The authors' method separates tracking from fine detail extraction using a video object segmentation (VOS) tracker to produce a temporally consistent coarse mask for each frame, while a dedicated matting pipeline refines boundaries:



Overview of the pipeline, where a flexible prompt and input video feed into a video object segmentation tracker to produce a coarse mask, which is refined by a Region-of-Interest (ROI) detector and converted into a trimap before a progressive multi-scale predictor generates the final high-detail matte.

A region-of-interest (ROI) detector then identifies regions with fine detail or semi-transparency, converting these into a trimap (a three-region mask dividing foreground, background, and uncertain areas) that guides refinement, after which a Progressive Alpha Predictor generates the final matte through a coarse-to-fine cascade across multiple scales

Conventional matting systems typically derive regions of interest using simple morphological operations, or by directly reusing the mask – approaches that can either overlook fine detail, or else include areas that do not require refinement:



Comparison of standard mask-based processing with ground-truth trimaps, showing how simple morphological operations produce coarse, uniform boundaries tha fine structures such as hair and transparency, leading to loss of detail in the final matte.

Compound Interest

Conversely, the proposed Region-of-Interest Detector instead treats this step as a pixel-wise classification task (i.e., treating each individual pixel in the image as a separate decision, and assigning it a label based on whether it belongs to a matting-critical region or not) that integrates the VOS mask, the current frame, and multi-scale image features, to more precisely isolate matting-critical regions.

In the new approach, the predicted ROI is first converted into a *pseudo*-trimap that separates definite foreground and background from uncertain regions, using the tracker mask to assign known areas while marking the ROI as ambiguous, so that subsequent processing can focus explicitly on boundaries where detail must be resolved.

Refinement is then handled by a Progressive Alpha Predictor that treats matting as a stepwise process, passing intermediate results from coarse to finer scales, with each stage using the image, the trimap, and the previous estimate to progressively sharpen structure and recover fine detail.

At the final stage, the highest-resolution output is upsampled to produce the completed matte, allowing broad shapes to be established early while finer elements such as hair and transparency are resolved in later passes.

During training, the authors froze the VOS tracker, while training *only the matting components* on high-quality image data – allowing fine detail to be refined without degrading tracking consistency. Supervision was then applied per frame, with regions of interest derived from the ground-truth alpha matte, and used to guide learning, while the ROI detector was trained with losses designed to encourage accurate boundary classification, and reduce jagged artifacts.

For alpha estimation, losses were applied across multiple scales to progressively improve detail, alongside an additional constraint intended to keep the predicted matte aligned with the original mask – helping to preserve structure, and to prevent hollow or broken regions in the final output.

Data and Tests

Eight image matting datasets were used initially, for the trials: I-HIM50K; P3M-10k; CelebAHairMask-HQ; AIM-500; Distinctions-646; AM-2K; UHRIM; and RefMatte; and three variants of the 'Sam2Matting' approach were developed as VOS trackers, respectively using SAM2.1-Tiny; SAM2.1-Base+; and the concept-focused SAM3.

The tracker component was frozen, with only the matting components optimized. All versions were trained for five epochs across four NVIDIA A6000 GPUs, each with a VRAM allocation of 48GB. A batch size of 32 was used, under the AdamW optimizer. Metrics used were Mean Absolute Difference (MAD); Mean Squared Error (MSE); Gradient (Grad); Connectivity (Conn); and dtSSD (for video matting only).

Quantitative Tests

The authors began with quantitative testing of the new systems on image matting, using the benchmarks P3M-500-NP; AM-2K ('GFM' in results); MAM ('Matte Anything', in results); E2E-HIM; Lightweight; and PPM-100 ('MODNet', in results):

	P3M-500-NP					AM-2K test					PPM-100				
Methods	MAD↓	MSE↓	Grad↓	Conn↓	SAD↓	MAD↓	MSE↓	Grad↓	Conn↓	SAD↓	MAD↓	MSE↓	Grad↓	Conn↓	SAD↓
P3M (Li et al., 2021a)	6.50	3.50	–	–	11.23	13.51	9.83	16.15	23.23	23.75	9.60	5.80	–	96.10	93.30
GFM (Li et al., 2022)	–	5.60	14.80	18.00	15.50	5.90	2.40	9.00	9.40	9.70	–	–	–	–	–
MODNet (Ke et al., 2022)	13.77	7.37	16.05	20.09	23.77	36.28	27.39	17.38	59.49	62.77	8.60	4.40	64.26	80.16	94.78
E2E-HIM (Liu et al., 2023)	5.40	3.00	–	–	9.25	–	–	–	–	–	7.20	4.00	–	–	–
MAM (Li et al., 2024b)	15.40	9.20	14.22	–	25.82	10.10	3.50	10.65	–	17.30	9.90	4.60	62.12	99.00	117.16
Lightweight (Zhong and Zharkov, 2024)	–	–	10.78	9.77	10.60	–	–	–	–	–	–	–	50.69	84.09	90.28
Matte Anything (Yao et al., 2024)	–	2.80	17.30	10.00	10.70	–	3.30	11.70	10.90	11.90	–	–	–	–	–
SAM2Matting (SAM2.1-T)	3.92	1.07	8.66	6.34	6.78	4.57	1.39	7.02	7.22	7.88	4.51	1.32	49.26	39.56	42.05
SAM2Matting (SAM2.1-B+)	3.81	1.00	8.78	5.97	6.58	4.90	1.59	7.23	7.75	8.44	4.31	1.22	49.17	37.58	40.27
SAM2Matting (SAM3)	3.83	0.97	8.48	5.84	6.61	4.32	1.19	6.75	6.61	7.43	4.23	1.16	47.55	36.27	39.45

Quantitative results on image matting benchmarks across P3M-500-NP, AM-2K test, and PPM-100, with lower values indicating better performance across all meti third-best results highlighted in red, orange, and yellow, respectively. Please refer to source paper for better resolution.

Of these results, the paper states:

> *'As shown [above], all three variants of SAM2Matting consistently outperform previous baselines across different metrics. For instance, the SAM2.1-Tiny variant achieves an 11.48 lower MAD than MAM on P3M-500-NP.'*

The authors maintain that the results across the board indicate that their approach achieves superior results through the core conceptual design, and not because of the level of data curation.

For video matting, SAM2Matting was evaluated on V-HIM60 and VideoMatte in a zero-shot setting, against the video-trained systems MatAnyone2, MatAnyone, MaGGIe, FTP-VM, and RVM, with consistent gains reported across both medium and hard splits.

Across all benchmarks, the three variants record lower errors on MAD, MSE, Grad, Conn, and dtSSD, with SAM3 achieving the strongest overall results, while the lowest dtSSD values apparently indicate more stable frame-to-frame consistency:

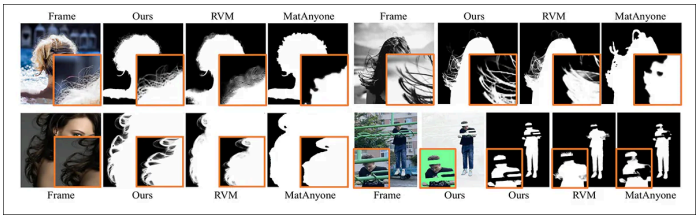
		V-HIM60-Medium					V-HIM60-Hard					VideoMatte-SD				
Methods	Venue	MAD↓	MSE↓	Grad↓	Conn↓	dtSSD↓	MAD↓	MSE↓	Grad↓	Conn↓	dtSSD↓	MAD↓	MSE↓	Grad↓	Conn↓	dtSSD↓
RVM (Lin et al., 2022)	[WACV'22]	–	–	–	–	–	–	–	–	–	–	6.08	1.47	0.88	0.41	1.36
InstMatt (Sun et al., 2022)	[CVPR'22]	19.34	–	7.21	6.02	24.98	27.24	–	7.88	8.02	31.89	–	–	–	–	–
FTP-VM (Huang and Lee, 2023)	[CVPR'23]	26.86	–	12.39	9.95	32.64	48.11	–	14.87	16.12	45.29	6.13	1.31	1.14	0.41	1.60
AdaM (Lin et al., 2023)	[CVPR'23]	–	–	–	–	–	–	–	–	–	–	5.30	0.78	0.72	0.30	1.33
SparseMat (Sun et al., 2023)	[CVPR'23]	18.20	–	8.03	6.87	30.19	24.83	–	8.47	8.19	36.92	–	–	–	–	–
MaGGIe (Huynh et al., 2024)	[CVPR'24]	13.85	–	6.31	5.11	23.63	21.23	–	7.08	6.89	29.90	5.49	0.60	0.57	0.31	1.39
MatAnyone (Yang et al., 2025b)	[CVPR'25]	29.95	19.72	9.03	12.28	5.98	30.09	18.87	8.93	10.00	6.72	5.15	0.93	0.67	0.26	1.18
MatAnyone2 (Yang et al., 2025a)	[CVPR'26]	15.12	5.86	6.36	5.43	4.50	45.75	35.03	8.43	14.75	6.16	4.73	0.55	0.51	0.19	1.12
SAM2Matting (SAM2.1-T)	–	13.76	4.61	7.78	5.01	4.23	18.58	8.79	8.03	6.16	5.37	4.85	0.41	0.36	0.20	1.22
SAM2Matting (SAM2.1-B+)	–	13.71	4.74	7.28	4.89	4.24	18.20	8.55	7.39	6.01	5.10	4.83	0.36	0.34	0.19	1.15
SAM2Matting (SAM3)	–	11.77	3.64	5.92	4.23	3.81	14.37	5.52	5.85	4.72	4.37	4.44	0.27	0.23	0.16	1.11

Quantitative results on video matting benchmarks across V-HIM60 and VideoMatte, evaluated in a zero-shot setting, showing lower errors across all metrics, with results highlighted in red, orange, and yellow respectively.

The authors contend that these outcomes reflect the decoupled design, where the VOS tracker preserves temporal structure and the matting modules focus on boundary detail, enabling image-trained models to exceed fully supervised video approaches.

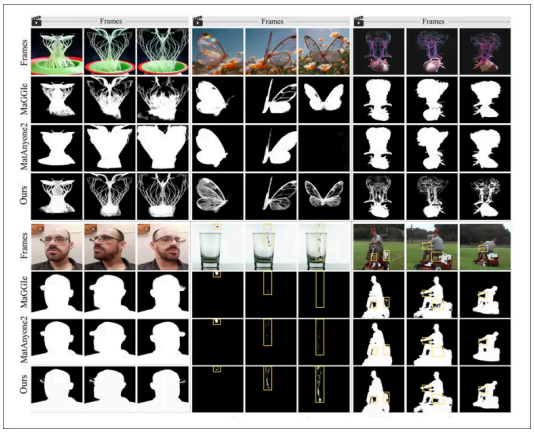
Qualitative Tests

For human matting in qualitative tests, the authors found that Sam2Matting outperformed competitive baselines:



Qualitative comparison on human and in-the-wild video matting, where SAM2Matting preserves fine hair strands and semi-transparent regions more accurately tha RVM and MatAnyone, producing cleaner boundaries and fewer missing structures in challenging areas.

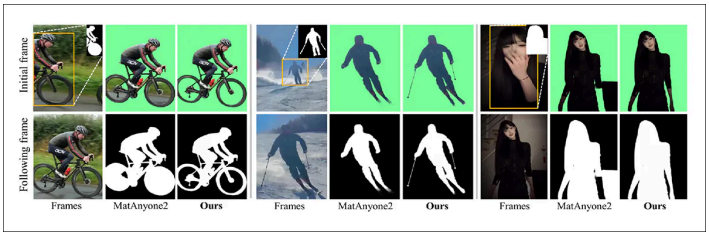
As shown below, existing video matting systems such as MatAnyone2 and MaGGIe, trained on domain-specific and often human-centric datasets, struggled to generalize to in-the-wild sequences, particularly when handling fast-moving subjects such as growing roots, semi-transparent butterflies, and rapidly dripping water:



Qualitative comparison on in-the-wild sequences, where SAM2Matting preserves fine structures and temporal consistency more effectively than MatAnyone2 and MaGGIe, particularly for non-human subjects, rapid motion, and semi-transparent elements such as water, glass, and insect wings. Please refer to source paper for better resolution.

Conversely, SAM2Matting proved able to maintain stable tracking, and to extract fine details more reliably, in these challenging scenarios.

Finally, as shown in the figure below, SAM2Matting effectively handled targets with attached objects, such as people riding bicycles or holding ski poles, while suppressing nearby background distractions, benefiting from the matte-mask consistency constraint described earlier.



Qualitative comparison on sequences with attached objects and background distractions, where SAM2Matting preserves structures such as bicycles and ski poles accurately than MatAnyone2 while suppressing nearby clutter and maintaining cleaner silhouettes across frames.

Conclusion

The achievements detailed in the new paper demonstrate the extent to which extraction, one of the oldest tasks in computer vision, remains unsolved and resistant to generalized approaches. As with many other vision-based models, the problem remains that extraction algorithms cling to domain knowledge instead of adapting easily to unseen and unknown objects; this particular task strains the outermost reaches of a model's generalization.

While the new work is a step forward from that dependence, there's still a long way to go, considering the extent to which this task is embedded in our daily lives, through video chat portals.

First published Monday, July 13, 2026

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