Training-Free Neural Matte Extraction for Visual Effects

Sharif Elcott*
Google
Japan
selcott@google.com

J.P. Lewis*
Google Research
USA
jplewis@google.com

Nori Kanazawa
Google Research
USA
kanazawa@google.com

Christoph Bregler
Google Research
USA
bregler@google.com

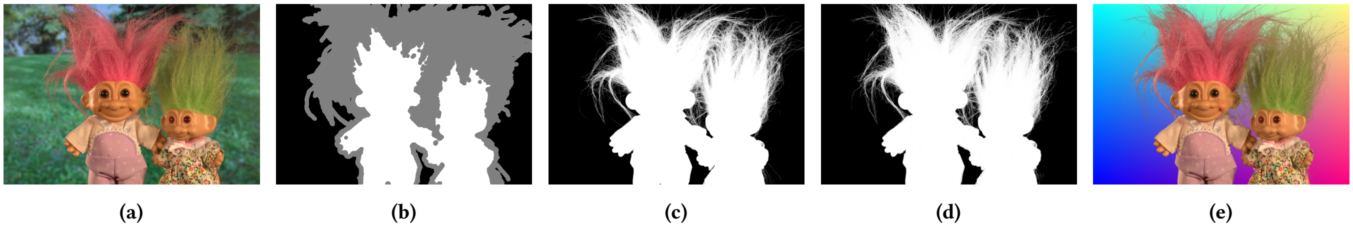

Figure 1: Given an image (a) and a crude "trimap" segmentation (b), we estimate a high quality alpha matte (d), allowing the foreground objects to be composited over arbitrary backgrounds (e). Image (c) shows the ground-truth alpha matte. Please enlarge to see details.

ABSTRACT

Alpha matting is widely used in video conferencing as well as in movies, television, and social media sites. Deep learning approaches to the matte extraction problem are well suited to video conferencing due to the consistent subject matter (front-facing humans), however training-based approaches are somewhat pointless for entertainment videos where varied subjects (spaceships, monsters, etc.) may appear only a few times in a single movie – if a method of creating ground truth for training exists, just use that method to produce the desired mattes. We introduce a *training-free* high quality neural matte extraction approach that specifically targets the assumptions of visual effects production. Our approach is based on the deep image prior, which optimizes a deep neural network to fit a single image, thereby providing a deep encoding of the particular image. We make use of the representations in the penultimate layer to interpolate coarse and incomplete "trimap" constraints. Videos processed with this approach are temporally consistent. The algorithm is both very simple and surprisingly effective.

CCS CONCEPTS

• **Computing methodologies → Image processing**; • **Applied computing → Media arts**.

KEYWORDS

Alpha matting, deep learning, visual effects.

[*] equal contribution.

SA '22 Technical Communications, December 6–9, 2022, Daegu, Republic of Korea
© 2022 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9465-9/22/12.
https://doi.org/10.1145/3550340.3564230

ACM Reference Format:
Sharif Elcott, J.P. Lewis, Nori Kanazawa, and Christoph Bregler. 2022. Training-Free Neural Matte Extraction for Visual Effects. In *SIGGRAPH Asia 2022 Technical Communications (SA '22 Technical Communications), December 6–9, 2022, Daegu, Republic of Korea.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3550340.3564230

1 INTRODUCTION

Alpha matte extraction refers to the underconstrained inverse problem of finding the unknown translucency or coverage α of a foreground object [Wang and Cohen 2007]. Matte extraction is widely used to provide alternate backgrounds for video meetings, as well as to produce visual effects (VFX) for movies, television, and social media. However, it is not always recognized in the research literature that these two applications have significantly different requirements. This paper introduces a neural matte extraction method specifically addressed to the assumptions and requirements of VFX.

Matting for video calls requires real-time performance and assumes a single class of subject matter, front-facing humans. This is a "train once and use many times" situation for which it is possible and advantageous to obtain training data. Video call matting often assumes a fixed camera, and may require a "clean plate" image of the room without the participant. On the contrary, VFX production has the following assumptions and requirements:

- **Diverse and often rare ("one-off") subject matter**. For example the youtube video [*Godzilla vs. Cat* 2021] (30M views) involves mattes of a cat, ships, and wreckage. A physical prop such as a crashed alien spaceship might only be used in a few seconds in a single movie, never to be seen again. Thus, **gathering a ground-truth training dataset for deep learning is often pointless**: if there is a method to generate the ground-truth mattes for training, just use this method to produce the desired mattes – no need to train a model.
- Visual effects frequently involves **moving cameras** as well as extreme diversity of moving backgrounds. For example,

an actor might be filmed as they run or travel in a vehicle any place on earth, or on set with extraterrestrial props in the near background. Thus, even for the major case of repeated subject matter (humans), gathering a representative training dataset is more challenging than in the case of video calls.

- Real-time performance is not required. Instead, the **on-set filming time is to be minimized,** due to the combined cost of the actors (sometimes with 7-figure salaries) and movie crew. It is often cheaper to employ an artist to work days to "fix it in post" rather than spend a few minutes on-set (with actors and crew waiting) to address an issue.
- **Clean-plates are not a desirable approach** for matting, and are often not possible. With moving cameras, a motion control rig is required to obtain a clean plate. The indoor use of motion control for clean plates is expensive due to the previous principle (minimizing on-set time). Moreover, this approach is generally not feasible outdoors, for reasons including background movement (e.g. plants moved by wind) and changing lighting (moving clouds blocking the sun).

Our solution, a matte extraction approach employing the deep image prior [Ulyanov et al. 2018], addresses these requirements. It has the following characteristics and contributions:

- It is a deep neural network matte extraction method targeted towards the requirements of VFX rather than video meetings. To our knowledge, this is the first deep neural matte extraction method that requires neither training data nor expensive on-set image capture to support the matting process (i.e. clean plates or greenscreens).
- It relies exclusively on the given image or video along with coarse "trimaps" (Fig. 1) that are easily created during post production using existing semi-automatic tools. Professional software such as Nuke or After Effects is typically used, however a familiar example is the "Select Subject" tool in Photoshop [Adobe 2018], followed by dilation/erosion and automated with "Actions".
- It does not require a clean plate, thus allowing extractions from outdoor footage.
- It does not require greenscreens, yet can produce detailed high quality mattes even when the foreground and background subject have similar colors (green hair in Fig. 1).

2 BACKGROUND AND RELATED WORK

The alpha-compositing equation is

$$I_i = \alpha F_i + (1 - \alpha) B_i \tag{1}$$

where $i \in \{r, g, b\}$, I_i are red, green, and blue values at a pixel of the given image (given), F_i are the foreground object color (unknown), B_i are the background object color (unknown), and $\alpha \in [0, 1]$ is the (partially unknown) "alpha matte", representing the translucency or partial coverage of the foreground object at the pixel.

Typical images contain large areas that are relatively easily estimated to be solid "foreground" ($\alpha = 1$) or "background" ($\alpha = 0$), as well as smaller regions where the alpha value is fractional. The matte extraction problem requires finding the these fractional alpha values. The problem is underconstrained because there are only three known values I_r, I_g, I_b but seven unknown values. Hair is the prototypical challenge for matte extraction due to its irregular coverage and translucency, though fractional alpha also generally appears along all object edges due to the partial coverage of a pixel by the foreground object. It is also necessary to estimate the unknown foreground color in the fractional alpha regions, since without this the foreground cannot be composited over a different background. Many methods assume that approximate demarcations of the solid foreground and background are provided, e.g. in the form of artist-provided "trimaps" [Rhemann et al. 2009] or scribbles [Levin et al. 2008].

While the research literature generally considers the problem of matte extraction from arbitrary natural backgrounds, in industry practice matte extraction from a greenscreen background is far from a solved problem and often requires an artist-curated combination of techniques to obtain the required quality [Erofeev et al. 2015; Heitman 2020]. LED walls remove the need for greenscreens in some cases, and are well suited for fast-paced television production, however they also have drawbacks: computationally expensive physical or character simulations are not possible due to the need for real-time playback in a game engine, bright backgrounds may introduce challenging light spill on the foreground physical objects [Seymour 2020], their cost is prohibitive for smaller studios, and the effects must be finished at the time of the principle shoot thus excluding the creative control and iterative improvement available in traditional post production.

Progress on matte extraction algorithms has been greatly facilitated by datasets that include ground truth (GT) mattes [Erofeev et al. 2015; Rhemann et al. 2009]. The GT mattes have been obtained by different means including chroma keying from a greenscreen and photographing a representative toy object in front of a background image on a monitor (Fig. 1 (a)). Failure to closely approximate these GT mattes indicates a poor algorithm, but an exact match may be unobtainable for several reasons: 1) the chroma key itself involves an imperfect algorithm [Erofeev et al. 2015], 2) light from multiple locations can physically scatter though (e.g.) translucent hair to arrive at a single pixel; this cannot be simulated in a purely 2D matte extraction process, 3) the image gamma or color space used in the benchmarks is not always evident.

The underconstrained nature of matte extraction has been approached with a variety of methods [Wang and Cohen 2007]. One classic approach estimates the unknown alpha at a pixel based on similarity to the distributions of known foreground and background colors in solid regions [He et al. 2011; Mishima 1992]. Another prominent principle finds the unknown alpha value by propagating from surrounding known values [Aksoy et al. 2018; Levin et al. 2008]. These approaches often require solving a system involving a generalized Laplacian formed from the pixel affinities in the unknown region, which prevents real-time or interactive use.

Deep learning (DL) approaches are used in recent research e.g. [Lin et al. 2020; Sun et al. 2021]. Much of this work focuses on providing alternate backgrounds for video meetings, and training databases consist of mostly forward-facing human heads. A state of the art method [Lin et al. 2020] demonstrates high-quality matte extraction on HD-resolution images at 60fps. Many DL methods adopt a combination of techniques, e.g. separate networks for overall segmentation and for fractional alpha regions, or other hybrid approaches.

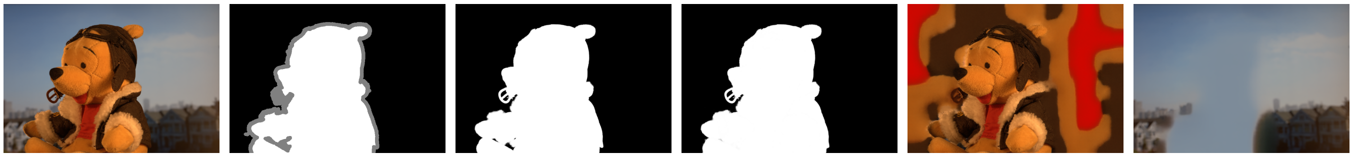

Figure 2: From left, image, trimap, GT alpha, estimated alpha, extrapolated foreground $\hat{F}$, extrapolated background $\hat{B}$.

2.1 Deep Image Prior

The Deep Image Prior (DIP) [Ulyanov et al. 2018] demonstrates that the architecture of an *untrained* convolutional network provides a surprisingly good prior for tasks such as image inpainting and denoising. The key observation is that while a powerful DNN can fit arbitrary image structures such as noise, it is "easier" to fit natural image structures, as reflected in faster loss curve decay. Most experiments [Ulyanov et al. 2018] optimize the weights of a U-net [Ronneberger et al. 2015] that maps *fixed* random noise to a *single* output image. Uncorrupted features of the image are fit earlier in the optimization process, so *early stopping* results in a corrected (denoised or inpainted) image.

The DIP has been successfully applied to other problems in image processing. Unsupervised coarse binary segmentation is demonstrated in [Gandelsman et al. 2019], based on a principle that it is easier (in terms of loss decay) to fit each component of a mixture of images with a separate DIP rather than using a single model. Concurrent with our work, [Xu et al. 2022] formulate a DIP approach to *background matting*. This problem scenario differs from our work in that it requires a clean plate, and hence is unsuitable for many VFX applications.

Our work also uses the DIP. We focus on high-quality estimation of fractional alpha mattes and, in contrast to previous work, undertake the challenging case where no clean plate is available, instead relying only on trimaps that can be easily created during post-production. The architecture and loss terms are detailed in the next section.

3 METHOD

Our method starts with a DIP network to reconstruct the target image. A second output head is added and tasked with inpainting the desired alpha in the trimap unknown region, constrained by the values in known regions. The main idea is that the first output head forces the network representation preceding it to "understand the structure of the image," while the second output head makes use of that information in estimating the matte.

In addition to the alpha output we add two additional networks which simultaneously reconstruct the foreground and background. Similarly to the first output of the first network, these networks' outputs are constrained to match the target image but, unlike the first network, they are constrained only in their respective regions of the trimap. Similarly to the alpha output, they extrapolate those constraints to inpaint the unconstrained region (see Fig. 2).

The latter three outputs – $\hat{\alpha}, \hat{F}, \hat{B}$ – are coupled via an additional constraint that they together satisfy (1) (see *Detailed Loss* below). We put $\hat{F}$ and $\hat{B}$ in separate networks (as in [Gandelsman et al. 2019]) rather than as additional outputs heads of the first network (as in [Sun et al. 2021]) in order to allow the inpainting of the foreground to be independent of the background. All three networks share the same generic U-net structure [Ulyanov et al. 2018], except for an additional output head in the first network. Our experiments use Adam [Kingma and Ba 2015] with a learning rate of 0.001.

3.1 Detailed Loss

The first term of our loss function is the reconstruction loss between the first network output and the target image:

$$L_I = \frac{1}{|I|} \sum_{i \in I} \|\hat{I}_i - I_i\|^2$$

The second loss term constrains $\hat{\alpha}$,

$$L_\alpha = \frac{1}{|C|} \sum_{i \in C} \|\hat{\alpha}_i - T_i\|^2$$

where T is the trimap and $C = F \cup B$ is the *constrained* region of the trimap.

The reconstruction losses for the foreground and background outputs are defined similarly to L_I, but constrained only in their respective regions of the trimap:

$$L_F = \frac{1}{|F|} \sum_{i \in F} \|\hat{F}_i - I_i\|^2 \qquad L_B = \frac{1}{|B|} \sum_{i \in B} \|\hat{B}_i - I_i\|^2$$

The three networks' outputs are coupled via the alpha-compositing equation as follows:

$$L_c = \frac{1}{|U|} \sum_{i \in U} \|I_i - (\hat{\alpha}_i \cdot \hat{F}_i + (1 - \hat{\alpha}_i) \cdot \hat{B}_i)\|^2$$

where $U = I - C$ is the *unconstrained* region of the trimap.

Finally, we include an exclusion loss [Sun et al. 2021] to prevent the structure of the foreground from leaking into the background and vice-versa:

$$L_e = \frac{1}{|U|} \sum_{i \in U} \|\nabla \hat{F}_i\|_1 \|\nabla \hat{B}_i\|_1 + \|\nabla \hat{\alpha}_i\|_1 \|\nabla \hat{B}_i\|_1$$

The total loss is the sum of the above six components:

$$L = L_I + L_\alpha + L_F + L_B + L_m + L_e$$

Unlike other DIP-based techniques, our algorithm does not require early stopping since the goal is to exactly fit both the image and the trimap constraints.

3.2 Temporal continuity

In our experiments temporal continuity was obtained by warm-starting the optimization with the final weight values of the previous frame, and stopping with a loss threshold rather than a fixed number of iterations. This simple strategy produces good results even on the relatively difficult case of hair. It also reduces the compute time by roughly an order of magnitude.

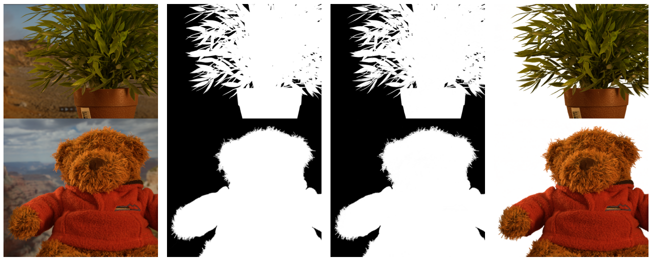

Figure 3: From left: image, ground truth alpha map, estimated alpha map, and composite over white. Please enlarge to see details.

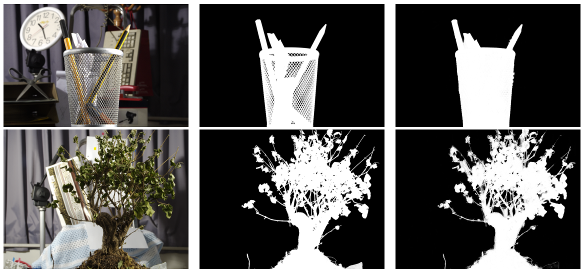

Figure 4: (Left) objects with holes can be a failure case. (Middle) ground truth alpha map. (Right) estimated alpha map.

4 RESULTS AND FAILURE CASES

We use images, trimaps, and GT alpha from the dataset [Rhemann et al. 2009] to allow comparison to the ground truth. Fig. 1 shows a moderately challenging case with extensive hair, some of which is similar to the background color. Fig. 3 shows several additional examples. While the ground truth and estimated alpha maps appear identical at first glance, small differences can be seen upon enlargement. These may be due to imperfections in our algorithm, but also may reflect differences in color spaces between our algorithm and the ground-truth estimation process.

Fig. 2 shows the extrapolated foreground $\hat{F}$ and background $\hat{B}$ for a particular image. Note that the inpainted colors need to be correct only in the unconstrained region surrounding the object border in order to allow compositing over a new background, as is the case in this figure. The unrealistic extrapolated $\hat{F}$ over the pure background regions is ignored.

4.1 Failure cases

While objects with "holes" sometimes yield good mattes (e.g. the plant in Fig. 3), they are also a failure case (Fig. 4). In some cases the trimap can be adjusted to highlight the missing holes, though this would be laborious in cases like the cup in Fig. 4.

5 LIMITATIONS AND CONCLUSION

We have introduced a matte extraction approach using the deep image prior. The algorithm is simple, requiring only a few tens of lines of code modification to an existing U-net. Our approach is training-free and is thus particularly suitable for the diverse, few-of-a-kind subjects in entertainment video production. It also may be of intrinsic theoretical interest in terms of the nature and solution of the matte extraction problem. A further potential use would be to produce ground-truth mattes to be used for DL training. As is the case with many matting algorithms, it assumes coarse guidance in the form of a trimap or similar constraints. This can be created by the artist using readily available semi-automatic tools.

Computational cost is the major limitation of the method, in common with classic methods [Levin et al. 2008]. Compute times for the examples shown in the paper are measured in minutes (but not hours) on a single previous generation Nvidia Volta GPU. This restricts the use of our algorithm to high-quality offline applications where extensive non-real-time computation is the norm, primarily movies and videos. On the other hand, the computation can take advantage of support for multiple GPUs provided in deep learning frameworks, and intermediate results can be visualized.

Our method can produce temporally consistent matte extractions from video by warm-starting the optimization from the previous frame (see accompanying video), however in our experience this requires that the trimaps have smooth motion from frame-to-frame. A topic for future work is to consider recurrent or other network architectures that might make the trimap choice more forgiving. This paper has focused on introducing the DIP matting algorithm. There was relatively little architecture and parameter exploration, and further improvements may be possible.

REFERENCES

Adobe 2018. How to use Select Subject in Photoshop for One-Click Selections. https://www.photoshopessentials.com.

Yagiz Aksoy, Tae-Hyun Oh, Sylvain Paris, Marc Pollefeys, and Wojciech Matusik. 2018. Semantic soft segmentation. *ACM Trans. Graph.* 37, 4 (2018).

Godzilla vs. Cat 2021. https://www.youtube.com/watch?v=nf7GsKFepDg.

Mikhail Erofeev, Yury Gitman, Dmitriy Vatolin, Alexey Fedorov, and Jue Wang. 2015. Perceptually Motivated Benchmark for Video Matting. In *British Machine Vision Conference.*

Yossi Gandelsman, Assaf Shocher, and Michal Irani. 2019. "Double-DIP": Unsupervised Image Decomposition via Coupled Deep-Image-Priors. In *Comp. Vision and Pattern Recognition.*

Kaiming He, Christoph Rhemann, Carsten Rother, Xiaoou Tang, and Jian Sun. 2011. A global sampling method for alpha matting. *Comp. Vision and Pattern Recognition.*

G.G. Heitman. 2020. Communication from technical artist, Weta Digital.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *Int. Conf. Learning Representations.*

Anat Levin, Dani Lischinski, and Yair Weiss. 2008. A Closed-Form Solution to Natural Image Matting. *IEEE Trans. PAMI* 30, 2 (2008).

Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. 2020. Real-Time High-Resolution Background Matting. *arXiv* (2020).

Y. A Mishima. 1992. Software Chromakeyer Using Polyhedric Slice. In *NICOGRAPH.*

Christoph Rhemann, Carsten Rother, Jue Wang, Margrit Gelautz, Pushmeet Kohli, and Pamela Rott. 2009. A Perceptually Motivated Online Benchmark for Image Matting. In *Comp. Vision and Pattern Recognition.*

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. 2015. U-Net: Convolutional Networks for Biomedical Image Segmentation. *MICCAI* (2015).

Mike Seymour. 2020. Art of LED Wall Virtual Production. fxguide.com.

Yanan Sun, Chi-Keung Tang, and Yu-Wing Tai. 2021. Semantic Image Matting. In *Comp. Vision and Pattern Recognition.*

Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. 2018. Deep Image Prior. In *Comp. Vision and Pattern Recognition.*

Jue Wang and Michael F. Cohen. 2007. Image and Video Matting: A Survey. *Found. Trends Comput. Graph. Vis.* 3, 2 (2007).

Yong Xu, Baoling Liu, Yuhui Quan, and Hui Ji. 2022. Unsupervised Deep Background Matting Using Deep Matte Prior. *IEEE Trans. Circuits Syst. Video Technol.* (2022).