Relighting Video Using Camera-Aligned Material Planes

Natali T. Chavez [1], Andrew S. Gordon [2]

[1] *Aristotle University of Thessaloniki, School of Fine Arts, Department of Film, Thessaloniki, 56430, Greece*

[2] *University of Southern California, Institute for Creative Technologies, Los Angeles, CA 90094, USA*

Abstract

In 3D animated films, Physically-Based Rendering (PBR) provides filmmakers with the opportunity to design and modify expressive and artistic lighting environments throughout the development process. However, in live-action filmmaking, contemporary visual effects (VFX) techniques allow for the addition and subtraction of entities in filmed performances, but provide only limited opportunities to substantially change the lighting environment in post-production. In this paper we describe a method for relighting a video of an actor's performance with a novel 3D lighting environment by combining AI-based image manipulation with traditional 3D visual effects techniques. Our approach is to segment foreground and background onto two different image planes positioned and scaled to fill a virtual camera's entire field of view as it moves in a 3D lighting environment, with distance-to-camera inferred using monocular metric depth models. Consistent lighting is achieved by inferring material properties of each video frame on each plane, e.g., normal maps, to enable physically-based rendering in 3D modeling software (Blender). We demonstrate that performances originally captured in diffuse lighting environments can be later relit in artistic ways, offering filmmakers a new option for low-cost virtual production.

Keywords

virtual production, post-production relighting, camera-aligned material planes, physically based rendering, open-source AI models, depth estimation, surface normals, virtual environment, Blender VFX pipeline, post-production lighting design

1. Introduction

When rendering virtual 3D environments as 2D images, photorealistic results require the careful modeling of geometry, camera intrinsics, light sources, and the material properties that dictate how light is reflected and transmitted when intersecting surfaces or traversing volumes. For makers of 3D animated films, the upside of this complexity is in the malleability of the resulting imagery, where radical changes to lighting and geometry can be readily made in software at any point in the production process in support of artistic expression. For live-action filmmakers, visual effects (VFX) afford some opportunities to replace foreground or background geometry in post-production, but few opportunities to substantially change the light that interacts with real-world actors and objects, beyond color grading. This limitation stems primarily from the technical infeasibility of capturing, at the time of filming, the full 3D geometry and material properties of a live-action scene, as would be needed for Physically-Based Rendering (PBR) under a novel lighting condition. However, recent advances in AI-based image analysis have made it possible to infer several aspects of a scene's geometry and its

[1]DCAC 2026: 8th International Conference on Digital Culture & AudioVisual Challenges, Interdisciplinary Creativity in Arts and Technology, Corfu-Greece & Online May 08–09, 2026
EMAIL: ntsavezn@film.auth.gr (A. 1); gordon@ict.usc.edu (A. 2);

.

material properties directly from photographs, e.g., the metric depth of each pixel from the camera, and the surface normal of each pixel. When these inferred properties of a 2D image are applied to plane mesh in a virtual 3D environment, the entities in the image can react to the virtual lighting environment and appear as if they were modeled as 3D objects. While there is substantial room for improvement in the accuracy of these image analysis models, as well as the breadth of material properties that can be inferred, they afford new opportunities for manipulating the lighting environment of live-action performances after they have been recorded on video. In this paper, we describe a new method for relighting live-action video that combines recent AI-based image analysis with traditional VFX techniques used in 3D computer graphics software, e.g., Blender. Our approach is to apply an assortment of models to infer geometric and material properties for each frame of a video recording, apply these properties to 2D mesh planes in a 3D virtual lighting environment, then align these material planes to a virtual camera such that the rendered images match the resolution of the original recording, but where each pixel is rendered in novel virtual lighting environments. We combine traditional VFX methods for camera tracking with inferred foreground and background depth to account for moving cameras and moving actors. After presenting our method, we describe its application in the virtual production of an experimental cinematic film, where performances captured in flat diffuse real-world lighting environments were rendered using PBR in novel virtual lighting environments. The main contribution of this work is in providing filmmakers a new option for low-cost virtual production that does not require special hardware or lighting equipment at the time of filming.

2. Method

We present our approach in three parts. First, we describe a concept of a Camera-Aligned Material Plane (CAMP) that positions a 2D video in a 3D environment. Second, we describe how the material properties of each frame in the video are inferred using an assortment of AI-based image analysis models. Third, we describe how traditional camera-tracking techniques for VFX are combined with inferred depth to position the planes correctly even when both the camera and foreground actors were in motion at the time of filming.

2.1. Camera-Aligned Material Planes

In 3D computer graphics software such as Blender, an image plane is a 2D rectangle positioned in 3D space, where the surface of the rectangle displays a photograph or other image. Such planes may assign a mask to an alpha channel that removes the background from the image, allowing a 2D foreground object to be placed anywhere in the 3D scene. A material plane sets the material properties of the image plane (e.g., diffuse, specular, normal layers) to enable Physically Based Rendering, allowing an object depicted on the plane to interact with the virtual lighting environment as if it were another 3D object in the scene. A Camera-Aligned Material Plane (CAMP) positions and rotates the plane to be flush and in the center of the virtual camera's viewport, with a virtual distance to the camera the same as the distance between the real-world world camera and the real-world subject, and the scale of the plane set to fill the virtual camera's entire field of view. When positioned and scaled correctly, a CAMP appearing in a virtual camera retains the pixel resolution of the original real-world imagery, but where each pixel of the foreground object has been relit using Physically Based Rendering. In this work, we create two separate CAMP objects in Blender from a contiguous video recording, representing the foreground and background of a cinematic performance. The properties of the two planes are identical except for the assignment of an alpha-channel mask to the foreground CAMP, such that the unmasked pixels of the foreground video completely hide the corresponding pixels of the background video, positioned behind it. To correctly position each CAMP such that it always fills the virtual camera's viewport, we create a Child-Of constraint for the plane to the camera

so that the center of the plane is always on the z-axis of the camera. We then add a driver for the x,y scale of the plane that computes a value based on the plane's distance (z) to the camera and the tangent of the camera's view angle. As a consequence, each CAMP's location, rotation, and scale is completely determined by the camera's six degrees of freedom plus the distance (z) to the camera. Figure 1 illustrates how these two frames would be positioned at three different times in a certain video. At each time point, the two planes are positioned and scaled to completely fill the viewport of the (moving) camera. When the virtual camera motion aligns with the trajectory of the real-world video camera and the depth value is animated correctly, the trajectory of the foreground plane in the virtual environment will align with the actual movement of the foreground actors in the real world. Accordingly, the two planes will react to light cast in different locations of the virtual 3D environment, which can be either static or dynamic based on the filmmakers' creative needs[2].

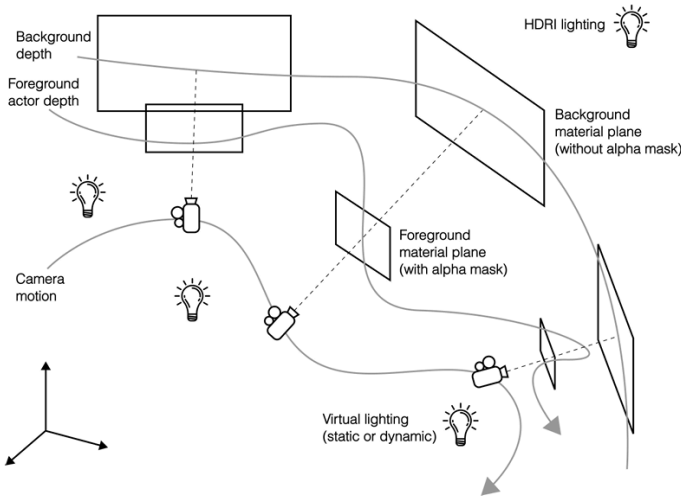

Figure 1: Camera-Aligned Material Planes with moving camera and moving foreground actor in a virtual lighting environment.

2.2. Inferring Material Properties

In our approach, the material properties of the CAMP are assigned using Blender's bidirectional scattering distribution function (BSDF) shader, which allows many surface properties to be encoded as separate images for each video frame, including base (diffuse) color, metallic degree, roughness, index of refraction, alpha mask, and surface normals. In this current work, we provide only the base color, alpha mask, and surface normals, leaving the remaining material properties to be explored in future research. Furthermore, we use the frames from the original unmodified video recording as base (diffuse) color of the material. As a consequence, our current approach works best when the original recording was shot in a diffuse lighting environment lacking hard shadows or excessive metallic surfaces. Figure 2 (a) provides an example frame from video captured during our experiments. To infer the alpha mask for each video frame, we use the open-weight InSPyReNet model [3], a high-performing salient object detection model that infers pixel-level annotations in high resolution images. Figure 2 (b) shows the result of applying this model to the original image, and is representative of the quality of foreground masking in our experiments. To infer the surface normals for pixels in each video frame, we use the open-weight IC-Light model [9], which uses a ControlNet latent diffusion architecture to apply a novel 2D lighting condition to an input image [8]. Using the method provided by these authors, we relight each video frame using four different lighting conditions (white light from bottom, top, left, and right), and compute the surface normal for each pixel based on

[2] We provide an open-source Blender add-on to aid in importing CAMP items. https://github.com/asgordon/CameraAlignedMaterialPlane

the relative illuminance from each condition. Figure 2 (c) shows the result of applying this technique to the original image. Computing surface normals in this manner produces high-quality results that are reasonably stable across successive video frames, but requires a significant amount of compute power owing to the four latent diffusion images required. In our experiments, processing each frame of video at 4K resolution took 42 seconds on a Mac Studio M2 Max with 96 GB unified memory.

2.3. Camera Tracking and Depth Estimation

Our approach fully derives the location, rotation, and scale of each plane from the six degrees of freedom of the camera location plus the distance between the camera and the plane. To achieve stable relighting of the CAMPs, it is necessary that these seven values are the same for the virtual camera as were for the real camera used when filming. To allow for both a moving camera and a moving foreground actor, these values were derived from the original video footage, and imported into Blender as keyframe animations of the corresponding values during rendering. For the camera's six values, we used Blender's built-in VFX tools for motion tracking, which solves for the 3D camera motion based on the movement of visual markers in the 2D video, as implemented in the underlying libmv library [4]. This component of our approach mirrors that of a traditional VFX pipeline for adding 3D objects into live-action footage, and requires both time and skill to achieve accurate results in the absence of hardware-based motion sensors or stages prepared with fiducial markers. To estimate the distances between the camera and the foreground and background planes, we experimented with the application of Apple's recent Depth Pro model, a monocular metric depth estimation model trained on a large mix of real-world and synthetic datasets [1]. To compute a single metric depth for foreground and background planes, we first applied the model to each frame of the original video footage, as shown in Figure 2 (d). We then segmented the results using the inferred alpha mask, as previously described. The mean values for the metric depth in foreground and background segments was selected as each plane's (z) distance to the camera. After computing these values for each frame, they were imported into Blender as keyframe animations for these values. We observed significant jitter in the positioning of the plane across subsequent frames using the raw inferred values. To alleviate this noticeable jitter, we applied the "1€ Filter" as a low-pass filter on each set of time-series values [2], yielding depth trajectories that were roughly as smooth as the solved camera motion.

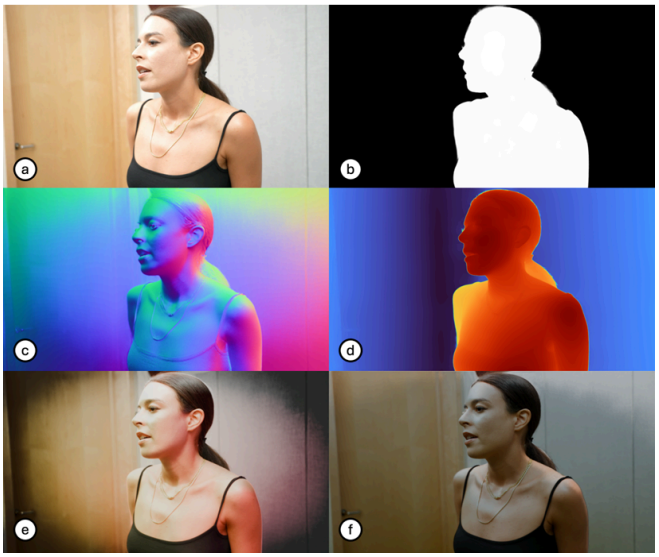

Figure 2: From top row, (a) original, (b) mask, (c) normal, (d) depth, (e) 3-point virtual lighting, (f) HDRI lighting.

3. Application to Virtual Film Production

The primary contribution of this research is in providing filmmakers a new alternative for low-cost virtual production that does not require special hardware or lighting equipment at the time of filming. As a proof-of-concept and to better understand the workflow required, we produced a short dramatic film that included a one-minute scene that was relit after filming using our method. In this scene, the protagonist is a singer rehearsing her vocal performance in a professional recording studio. Alone in the dimly-lit studio and nervous about her performance, she warms up her voice with various humorous singing exercises, and her nervousness gradually gives way to pent-up laughter. As a set location, we selected an empty interior office that had previously been used as a recording space for voice-over acting, with acoustic panels on the walls as might be seen in professional music studios. Rather than using the existing fluorescent tube lights in this office, we created a flat, diffuse lighting environment using two 300-watt daylight-balanced point-source LED lamps, each directed toward the ceiling. When shooting in this small space, our minimal production crew consisted of the director, a camera operator, a microphone boom operator, and the actress who played the musician in this scene. Scene footage was filmed at 24 fps using a 4K full-frame digital camera with a wide angle lens, mounted on a handheld 3-axis camera stabilizer. After filming, we selected 12 minutes of scene footage (5 continuous shots) for relighting using our method, which was applied before providing the material to the film editor who selected shots for the final one-minute scene. We found there were three time-intensive tasks in the application of our relighting method to this source footage. First, with over 17,000 frames in this footage, inferring the material properties and foreground depth required several hundred hours of compute time to complete. Second, tracking the camera motion in each of the 5 continuous shots took several days of skilled labor using Blender's built-in motion tracking tools, which was complicated by the unfortunate scarcity of trackable features on the walls of our selected set location. Third, after importing the final CAMPs into Blender, considerable effort and experimentation was directed toward designing a synthetic lighting environment that supported the artistic vision of the filmmaking team. Figure 2 (a) provides an example frame as captured during filming, which was used as the diffuse material layer when relighting. Figure 2 (e) is the corresponding frame rendered using a traditional 3-point lighting setup, with synthetic area lights for key, fill, and overhead lights of different colors. Figure 2 (f) is the same frame rendered using an HDRI background surface as a lighting environment. In these images and in our experiments, we used Blender's faster EEVEE renderer (version 4.2), which facilitated rapid experimentation with different virtual lighting options. We felt that the most artistically satisfying results were achieved by combining both an HDRI lighting environment with virtual area lights positioned to accent important beats of the actress's performance.

Table 1 quantifies the performance of our selected image processing models on the 24 fps 4K video captured for this production (processing time and estimated failure rate). Failure rate was estimated by randomly sampling 200 frames from the edited video and visually judging whether the frame would require substantial manual editing to avoid the introduction of distracting visual artifacts in the rendering of the material plane.

Table 1
Processing time (seconds/frame) and estimated failure rate for image processing models.

Image processing model	Mean seconds/frame	Estimated failure rate
InSPyReNet (background mask)	1.79	11.0%
IC-Light (surface normals)	42.02	4.5%

4. Limitations

There are several notable limitations of our approach that could be addressed through future technical advancements. First, each of the image processing models that we use in this research operate on single frames only, rather than video, causing noticeable discontinuities between frames in several cases. Our chosen foreground segmentation model (InSPyReNet) performed inconsistently across frames on hoop earrings, for example, causing a distracting effect where they sporadically appear as foreground disks rather than loops. Our chosen normal map model (IC-Light) performed inconsistently across frames for the eyes of actors, which was distracting under certain lighting conditions. Predictions of background depth (Apple Depth Pro) were inconsistent across frames when filming closeups of an actor's face or when backgrounds lacked distinguishing features, even after applying a low-pass filter. In addition to improved accuracy, future applications of our method would greatly benefit from models that aimed for consistency across image sequences. Second, there are additional material properties of images that should be inferred to improve the photorealism of the rendered scenes. Our current approach uses the original imagery as the diffuse layer of the CAMP, but we expect that better results could be obtained by inferring the roughness, specularity, and subsurface scattering properties separately, in combination with an inferred diffuse albedo layer. Our approach is well-suited to emerging work on rendering with neural material shaders, provided that properties could be inferred directly from input image sequences [6][7]. Third, there are some peculiarities of our approach resulting from the use of planes as replacements for foreground 3D geometry. For example, virtual lights positioned behind a foreground plane do not cast direct light on the plane's front surface, and must be repositioned in order to cast light on shoulders and hairlines, for example. The shadows cast by foreground planes can also be bizarre under certain lighting conditions. A spotlight from one side of a foreground plane may only cast a slim shadow on a surface on the opposite side. Likewise, shadows cast on surfaces behind the foreground plane can reveal that its span is limited to what is present in the camera's frame, e.g., the shadow cast by an actor in a medium shot might abruptly end where it should be continued by the actor's off-camera lower body. Finally, our plane-based approach fails when the foreground subject is oriented such that it spans a long depth of field, most notably when there are two or more actors in the foreground plane at different camera depths. In such cases, it would be necessary to expand the approach to allow for multiple actors to appear on separate material planes.

5. Discussion

The relighting approach described in this paper combines traditional VFX methods and software with emerging AI-based image analysis models, enabling live-action filmmakers an ability to design and modify the lighting environment of filmed performances in post-production. The three essential inputs to this approach, namely the original footage, the lighting environment, and the camera motion, are each processed and then integrated within traditional 3D graphics software (Blender), where CAMPs are rendered using Physically-Based Rendering. However, we foresee that continued research progress on diffusion-based image manipulation may one day obviate the need to render frames within these software tools. 2D image relighting models such as IC-Light, which we employed when computing surface normals in our current work, might in the future be conditioned on both the camera pose and the neural radiance field (NeRF) of a captured (real) or designed (synthetic) 3D lighting environment. In the shorter term, our future work focuses on removing from our approach the laborious step of tracking camera motion in Blender for each clip using hand-annotated markers. We expect that our method is well-suited for the application of Hierarchical Localization methods [5] that combine Structure-From-Motion and image retrieval to determine camera motion, especially where moving foreground objects (actors) can be automatically segmented out of each frame. As well, we see opportunities to integrate hardware-based camera tracking into our production pipeline, including both the inside-out and outside-in methods used by current Virtual Reality headsets. Likewise, existing professional motion-capture studios would be well-positioned to integrate our approach,

which could further expand the utility of these spaces as virtual production stages for live-action filmmaking.

6. Acknowledgements

The project or effort depicted was or is sponsored by the U.S. Army Combat Capabilities Development Command under contract number W912CG-24-D-0001. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred

7. References

[1] A. Bochkovskii, A. Delaunoy, H. Germain, M. Santos, Y. Zhou, S. R. Richter, and V. Koltun. Depth Pro: Sharp monocular metric depth in less than a second. arXiv preprint arXiv:2410.02073, 2024. URL: https://arxiv.org/abs/2410.02073.

[2] G. Casiez, N. Roussel, and D. Vogel. 1€ filter: A simple speed-based low-pass filter for noisy input in interactive systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*, 2012, pp. 2527–2530.

[3] T. Kim, K. Kim, J. Lee, D. Cha, J. Lee, and D. Kim. Revisiting image pyramid structure for high resolution salient object detection. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*, 2022, pp. 108–124.

[4] K. Mierle. Open source 3D reconstruction from video. Master's thesis, University of Toronto, 2008.

[5] P.-E. Sarlin, C. Cadena, R. Siegwart, and M. Dymczyk. From coarse to fine: Robust hierarchical localization at large scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[6] P. Kocsis, L. Höllein, and M. Nießner. IntrinsiX: High-quality PBR generation using image priors. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2504.01008. URL: https://arxiv.org/abs/2504.01008.

[7] N. Raghavan, K. Mullia, A. Trevithick, F. Luan, M. Hašan, and R. Ramamoorthi. Generative neural materials. In *ACM SIGGRAPH 2025 Conference Papers (SIGGRAPH '25)*, Article 162, pp. 1–11. ACM, 2025. https://doi.org/10.1145/3721238.3730746.

[8] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2023, pp. 3836–3847.

[9] L. Zhang, A. Rao, and M. Agrawala. IC-Light GitHub repository, 2024. URL: https://github.com/lllyasviel/IC-Light.