Nature Unveiled: Recovering Occluded Landscape from Video Sequence

Anonymous CVPR submission

Paper ID *****

Abstract

Background recovery in videos of natural scenes is highly desirable given the increased movement and clutter of human activity in natural spaces. This study aims to remove occlusion and recover the natural backdrop from a video sequence. The process begins by segmenting the video frames into foreground and background. Among various segmentation methods tested, using learned optical flow between frames yielded the best results. After calculating a segmentation/mask for each frame, the next step is to inpaint the foreground. Inpainting methods tested include Stable Diffusion Inpainting and Flow-edge Guided Video Completion (FGVC). While Stable Diffusion Inpainting produced the highest quality frames, it ultimately lacked consistency between frames. The discussion includes potential future work to ensure frame-to-frame consistency.

1. Introduction

Background recovery in videos of natural scenes is highly desirable due to the increased movement and clutter of human activity in these spaces. Objects such as telephone lines, fences, houses, and trees can obscure views of scenic landscapes. This project aims to recover the background and remove occlusions from a video sequence shot from the passenger side of a car, leveraging information from neighboring frames to "see" behind occlusions.

1.1. Previous Work

Previous work on background recovery from video includes several approaches. For instance, [8] focuses on using the estimated motion of moving occlusions to recover background scenes. However, this method assumes a stationary camera and relies on multiple frames where the occluded object is moving to reconstruct the background behind it.

In [9], the authors leverage motion parallax from slight camera movements and accompanying optical flow fields to separate foreground and background. This approach successfully removes reflections and fence occlusions through a unified framework, demonstrating the potential of combining slight camera motion with advanced optical flow techniques.

Another significant contribution is from [7], where the authors aim to remove occluded objects such as cars, pillars, and people to clean up photos for Image-Based Rendering (IBR). They use multiple images to create a rough 3D rendering of a scene and then employ this 3D model to inpaint masked regions, thereby enhancing the visual quality and coherence of the reconstructed scenes.

This work builds on these approaches by focusing on dynamic scenes with a moving camera, aiming to improve frame consistency and the quality of the recovered background. Future work will explore methods to ensure temporal consistency between frames, addressing the challenges identified with current inpainting techniques.

Here is a revised and expanded version of the Problem Statement section:

2. Problem Statement

Assume that there is a background B that is occluded by a foreground F. We are given multiple continuous frames in a video where we can estimate the depth to separate the objects in B and F. Due to the changing position of the camera and the non-uniform and semi-transparent nature of F in some areas, we can accumulate information about B over time.

The goal is to use this accumulated information to inpaint the photos in a manner that is consistent across views. We aim to compare this multi-frame inpainting method with single-frame inpainting results to evaluate improvements in consistency and quality.

2.1. Dataset

Finding an appropriate synthetic dataset with ground truth data for this task proved challenging. Most available landscape or driving footage, both synthetic and real, is typically captured with the camera mounted on the front of the vehicle rather than to the side, which is necessary for our specific occlusion recovery problem.

Currently, I am testing my method on a video sequence that I captured. This sequence is 4 seconds long and consists of 120 frames moving from right to left. Frames 0 and 1 are shown in Figure 1.

(a) Frame 1 (b) Frame 0

Figure 1. Video frames with occluded landscape. Notice the camera is moving to the left.

The dataset for this project is a custom video sequence shot from the passenger side of a car. This allows for dynamic occlusion scenarios where the camera moves and captures varying perspectives of the occluded background. The sequence captures the complexity of real-world occlusions and provides a challenging testbed for evaluating the proposed background recovery method.

3. Technical Approach

To solve the problem of background recovery in videos with occlusions and moving cameras, the following technical approach will be employed:

3.1. Segmentation

The first step is to segment the frames of the video into foreground (F) and background (B). Accurate segmentation is crucial for isolating the regions that need inpainting. The segmentation process relies on caluclating a depth map for each frame. This can be done using:

- **Stereo Depth Map (neighboring frames):** Utilizing the disparity between consecutive frames to estimate object depth.
- **Monocular Depth Estimation (single frame):** Using deep learning models trained on large datasets to predict depth from a single frame.
- **Learned Optical Flow:** Leveraging parallax effect (objects far away move less across frames) and using optical flow between frames as an indicator of object depth.

3.2. Masked Inpainting

Once the foreground regions are identified and masked, inpainting techniques are applied to recover the background:

- **Text-guided Inpainting with Stable Diffusion:** Utilizing Stable Diffusion models to inpaint masked regions guided by text prompts. This method focuses on generating high-quality images but needs to address temporal consistency across frames.
- **Flow-edge Guided Video Completion (FGVC):** Using optical flow and edge information to guide the inpainting process, ensuring that the inpainted regions are temporally coherent and consistent with the video structure.

3.3. Comparison of Methods

The performance of the inpainting methods will be compared based on several criteria:

- **Visual Quality:** Evaluating the perceptual quality of the inpainted frames for naturalness and realism.
- **Consistency:** Assessing the temporal consistency across frames to avoid flickering and artifacts.

4. Segmentation Results

4.1. Stereo Depth Map Estimation

I experimented with stereo methods using OpenCV [1]. Given that the camera is moving along a straight trajectory, the resulting frames are nearly parallel, which facilitates the use of stereo vision techniques. The depth map generated from the disparities between frame 0 and frame 1 is shown in Figure 2, along with the corresponding segmented foreground and background images in Figure 3

For the trees and fence, the depth map performs quite well, accurately distinguishing these elements. However, for the sidewalk, grass, and curb, the uniform texture results in an incorrect depth map in those regions. This highlights a common challenge in stereo depth estimation where regions with uniform texture or repetitive patterns can lead to ambiguous disparity calculations.

4.1.1 Find Corresponding Points

I wondered if the depth map from stereo would be improved by rectifying the images since the images are not parallel. I used OpenCV [1] to find SIFT (Scale-Invariant Feature Transform) points and brute force block match 150 points on the image. The corresponding points and SIFT points are shown in Figure 4.

4.1.2 Rectify the Images w/ Corresponding Points

Using HW 2 Problem 2, I calcualted the homographies to create parallel images as seen in Figure 5. However, now the parallel images are different dimensions and OpenCV function only take in images of the same dimensions.

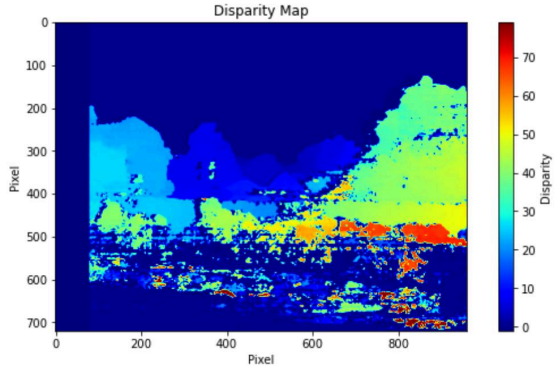
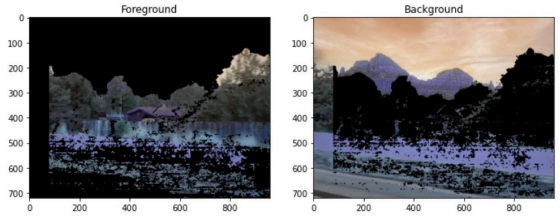


Figure 2. OpenCV Stereo calculated depth map from frame 0 and frame 1.



(a) Segmented Foreground (b) Segmented Background

Figure 3. Foreground and background segmented with stereo depth map.

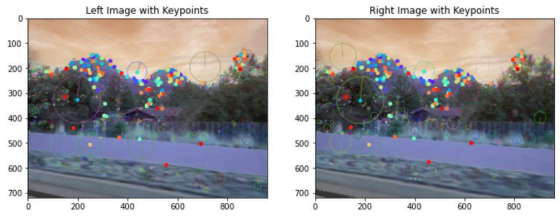


Figure 4. Corresponding points and SIFT points.

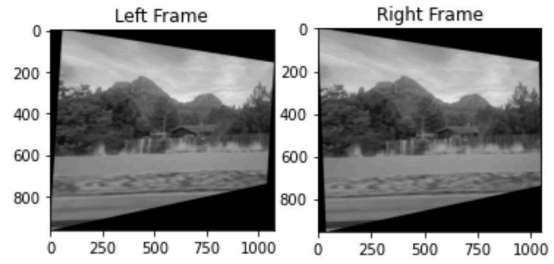


Figure 5. Rectified parallel images of frame 0 and frame 1.

4.2. Monocular Depth Estimation

Given the difficulties of stereo depth estimation, I instead turned to estimating depth from a single frame.

4.2.1 GLPN fine-tuned on NYUv2

I used a pretrained global-local path networks (GLPN) model [3] and the predicted depth map for Frame 0 is seen in Figure 6. The depth map is too smooth and notice the tree (circled in red) is blended out of the depth map. The resulting foreground and background segmentation is shown in Figure 7. Notice the house in the middle and trees behind it are not segmented correctly.

4.2.2 MiDaS

I tried another pretrained model MiDaS [4]. The estimated depth map of Frame 0 is shown in Figure 8 and corresponding foreground and background segmentation is shown in Figure 9. The model was successful in predicting the house in the middle to be on the same depth plane. However this depth map is too smooth to segment the trees in the middle into the foreground.

The subpar performance of monocular depth estimation led me to look at learned optical flow methods instead.

4.3. Learned Optical Flow

I used a pretrained optical flow model RAFT [6] in the Pytorch library. The optical flow for each frame was calculated with its following frame. Since objects that are closer to the moving camera will move at a faster speed than objects away from the camera, we can use the magnitude of the optical flow as an indicator of objects' depth in the scene. A random selection of frames with corresponding optical flow magnitude maps are shown in Fig. 10.

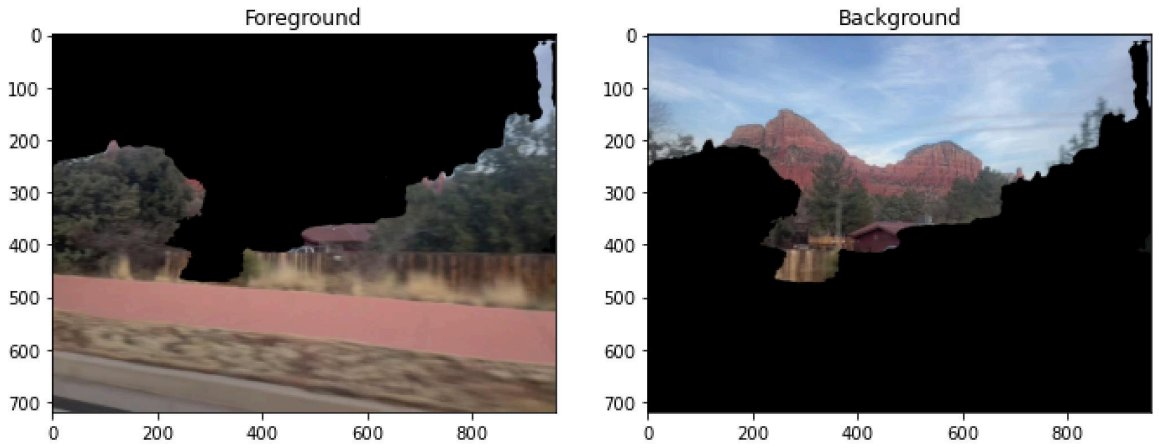
From the magnitude maps of each frame, I then looked at the best threshold to use for the foreground mask. An example of various threshold levels (relative to max value of optical flow magnitude of a frame) are shown in Fig. 11. The final masks that were used for inpainting were magnitude greater than 5 % of the max value of the optical flow magnitude.

5. Inpainting Results

5.1. Stable Diffusion Inpainting

Using a pretrained model from Hugging Face that is based off of Stable Diffusion that takes in a text prompt, image, corresponding mask and returns an image with the mask region inpainted according to the text prompt [5]. An example of outputs to the prompt "mountains in the desert" is shown in Fig. 12.

Figure 6. GLPN Monocular Depth Map on Frame 0



(a) Segmented Foreground

(b) Segmented Background

Figure 7. Foreground and background segmented with GLPN network.

Since the pretrained model processed each frame independently from another the resulting frames from a video were not consistent despite using the same text-prompts and random seeds as shown in Fig. 13.

5.2. Flow-edge Guided Video Completion

I then fed frames and corresponding masks into a flow-based video completion algorithm called "Flow-edge Guided Video Completion" (FGVC) [2]. This algorithm relies on optical flow to fill in masked regions across frames. An resulting frame is shown in Fig. 14. This method is normally used with mask of smaller area (e.g. removing a tennis player in a tennis match) and fails to find the correct information to fill in the large mask region of the frames.

6. Conclusion

Among the methods tested, optical flow proved to be the most effective for depth level segmentation compared to monocular and stereo depth estimation. It provided the most

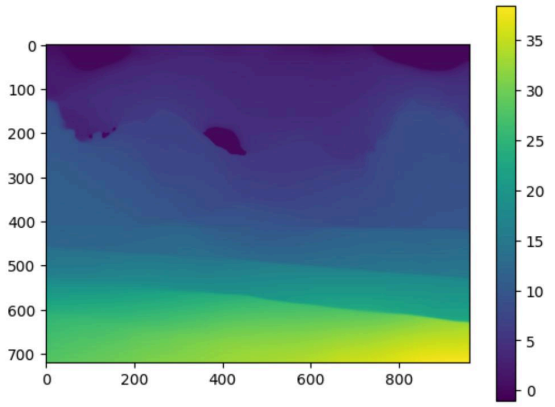


Figure 8. MiDaS Monocular Depth Map on Frame 0

accurate distinction between foreground and background objects.

However, inpainting using diffusion models, while capable of producing high-quality frames, requires specific prompts and lacks consistency across all frames even when using the same prompt. The Flow-edge Guided Video Completion (FGVC) method also fell short, as the regions to be inpainted were too large and there wasn't sufficient information across frames to accurately complete the masked regions.

To address the issue of inconsistent inpainting, a potential solution could involve training a Low-Rank Adaptation (LORA) on the video or fine-tuning the existing diffusion inpainting model. This approach would allow the model to take a prior image into account and penalize deviations from it, potentially improving temporal coherence and consistency in the inpainting results.

Future work will focus on these enhancements to improve the robustness and reliability of background recovery in video sequences.

References

[1] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000. 2

[2] Chen Gao, Ayush Saraf, Jia-Bin Huang, and Johannes Kopf. Flow-edge guided video completion. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XII 16*, pages 713–729. Springer, 2020. 4

[3] Doyeon Kim, Woonghyun Ga, Pyunghwan Ahn, Donggyu Joo, Sehwan Chun, and Junmo Kim. Global-local path networks for monocular depth estimation with vertical cutdepth. *CoRR*, abs/2201.07436, 2022. 3

[4] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer, 2020. 3

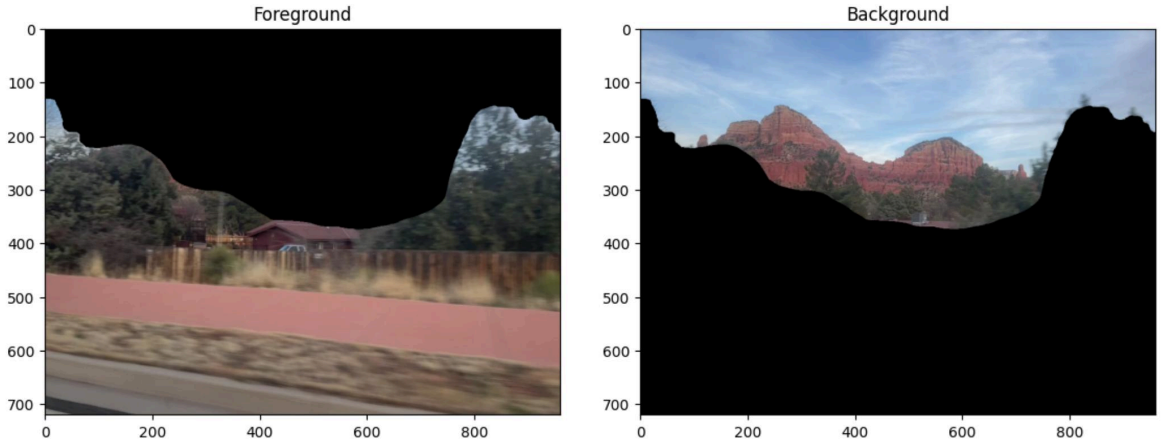
[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022. 3

[6] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 402–419. Springer, 2020. 3

[7] Theo Thonat, Eli Shechtman, Sylvain Paris, and George Drettakis. Multi-view inpainting for image-based scene editing and rendering. In *2016 Fourth International Conference on 3D Vision (3DV)*, pages 351–359. IEEE. 1

[8] Srenivas Varadarajan, Lina J. Karam, and Dinei Florencio. Background recovery from video sequences using motion parameters. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 989–992. ISSN: 2379-190X. 1

[9] Tianfan Xue, Michael Rubinstein, Ce Liu, and William T. Freeman. A computational approach for obstruction-free photography. 34(4):1–11. 1



(a) Segmented Foreground (b) Segmented Background

Figure 9. Foreground and background segmented with MiDaS network.

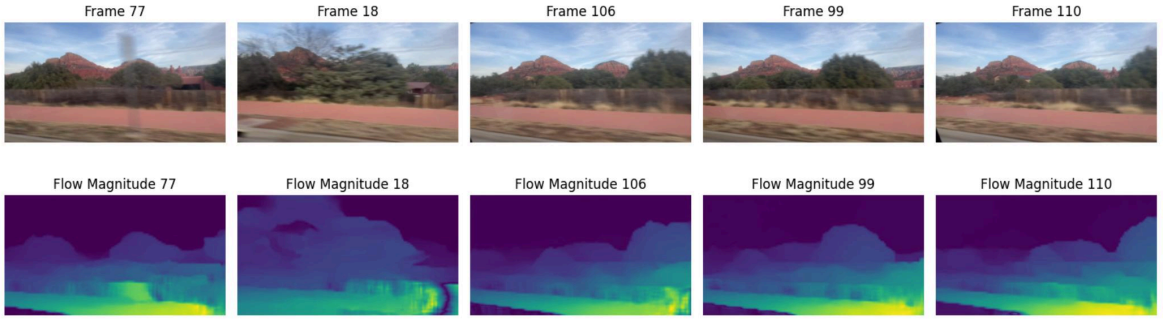


Figure 10. Optical Flow Magnitude for Random Frames from RAFT

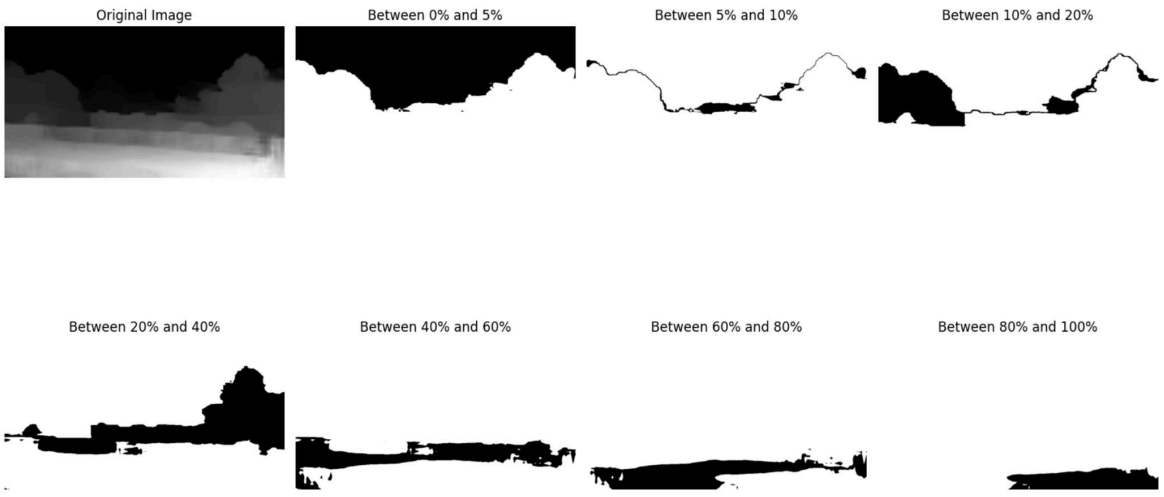


Figure 11. Example of Threshold Layers of Optical Flow Magnitude

Figure 12. Diffusion model outputs of masked image with prompt ”mountains in the desert”

Figure 13. Selection of Diffusion Model Inpainted Frames with ”desert landscape” text prompt

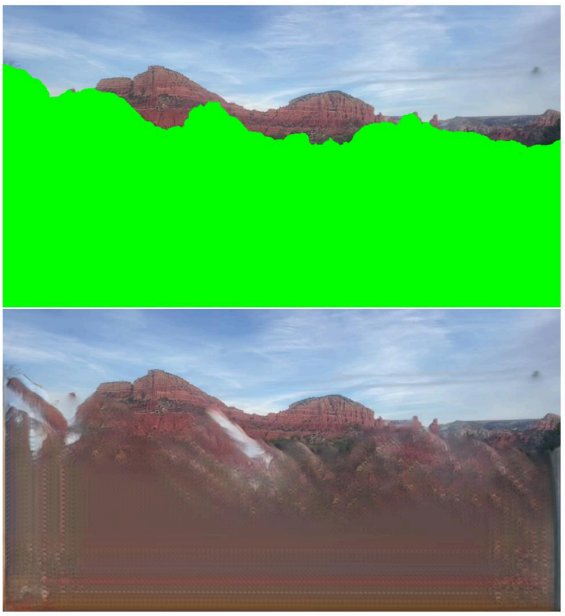


Figure 14. Image with mask applied and corresponding FGVC calculated frame