Cleaner StreetView

Dasheng Bi
Department of Electrical Engineering and Computer Sciences
University of California, Berkeley
dbi@berkeley.edu

Abstract

For sequences of photos acquired over a spatiotemporal range, such as those in StreetView, we often care more about static components and landmarks in the scene, whereas dynamic, moving objects may be distracting and produce artifacts when performing reconstruction or interpolation on the scene. Here, we propose a pipeline to obtain ***Cleaner StreetView*** *images, that combines inference on large, pretrained image models and classical stereo pose estimation techniques. We show several examples for visualization and highlight potential points for improvement.*

1. Introduction

With the increasing availability of large-scale image datasets captured over space and time ("StreetView" datasets) [1, 6], creating navigable virtual environments has become more and more accessible [5]. These datasets contain rich visual information about their corresponding environments. However, the dynamic nature of such environments poses challenges when attempting to reconstruct scenes or interpolate between frames. Moving objects, such as vehicles and pedestrians, can introduce inconsistencies and artifacts, that lower the quality of visualizations and clutter the scene, making it difficult for viewers to focus on the static components (Fig. 1).

There exists previous work that attempts to address this problem of disentangling moving objects from StreetView scenes. However, each method has its own limitations, such as only being able to handle pedestrians [3], only handle grayscale images related by homography transformations [7], or require depth information in input images [9].

Here, we propose a pipeline for obtaining cleaner StreetView images that fully utilizes the power of modern computer vision methods, including state-of-the-art pretrained image segmentation and inpainting models [2, 8]. We note that our pipeline is modular and these models can be readily replaced with other image models as the user desires. A schematic of the pipeline is available in Fig. 2.

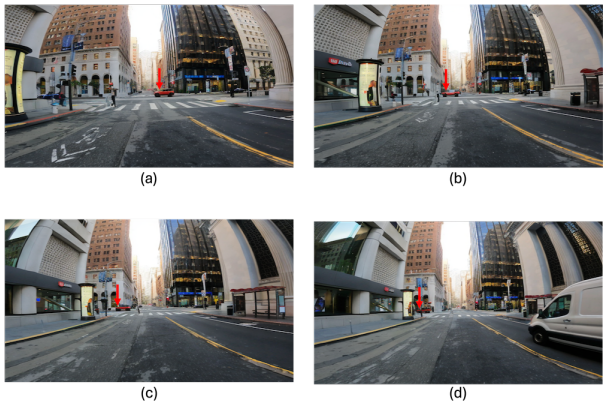

Figure 1. StreetView contains moving objects. (a)-(d) depict several StreetView frames, and the car pointed to by the red arrow is at a different relative location in the scene in each of the frames.

In particular, our pipeline takes as input a StreetView image sequence, namely a sequence of monocular RGB images of a scene with unidentified camera poses and camera parameters. First, we segment the images into individual objects using a large, pretrained panoptic segmentation model [2]. This yields a set of object masks for each frame. Then, we classify each object as static or moving based on the agreement of keypoints with an estimated global fundamental matrix between the current and previous frame. Objects with high reprojection loss are considered moving, and are subsequently removed. Finally, we use a pretrained diffusion model to inpaint the gaps created by removing the identified moving objects [8].

This project utilizes recent advancements in computer vision to improve the quality of StreetView images, which has many potential downstream applications such as providing better reference street scenes, as well as enabling 3D street reconstructions with less artifacts.

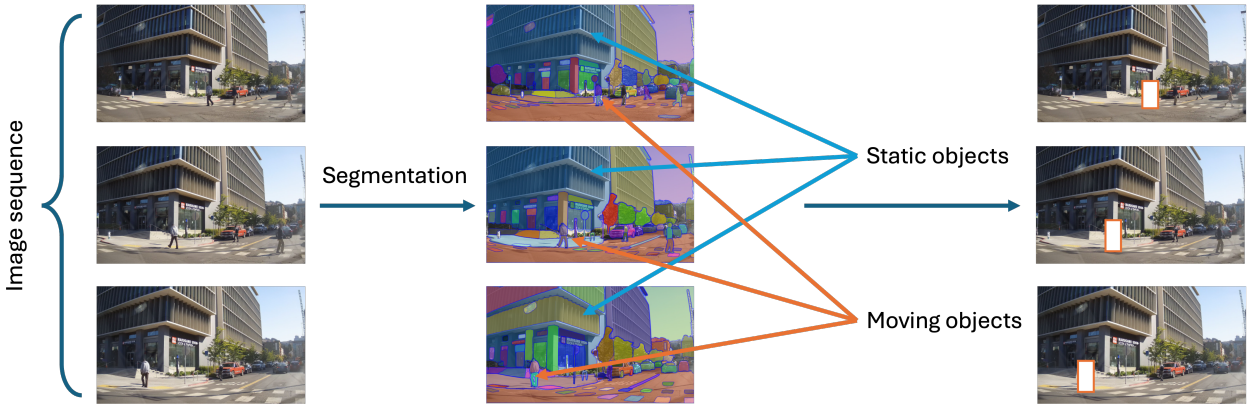

Figure 2. Schematic of our pipeline. The input is a StreetView image sequence consisting of multiple frames of a scene from various unidentified camera poses. We perform frame-wise panoptic segmentation to obtain object masks. Then, for each object in each frame, we classify it as moving or static. Finally, for objects that we identify as moving, we remove them using the object mask and inpaint the frame using a diffusion model.

2. Results

2.1. Panoptic segmentation

To identify individual objects in the scene, we simply pass each frame of the StreetView image sequence through a pretrained image model trained on panoptic segmentation. Here, we use Mask2Former [2], but other segmentation models can be readily used as well. As shown in Fig. 3, the model is able to reliably segment out each object present in the image (in this case, the moving car), even as the object becomes more distant and is partially occluded in later frames.

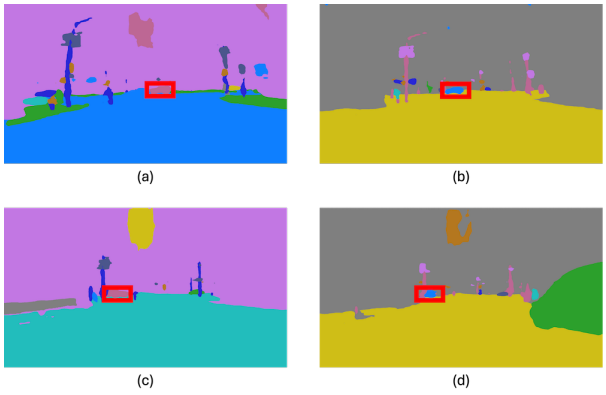

Figure 3. Results of frame-wise panoptic segmentation for a StreetView image sequence. (a)-(d) depict segmentation masks for several StreetView frames, and the car in Fig. 1 is reliably segmented out by the model (see red bounding boxes).

2.2. Moving object identification

The key step of the pipeline is identifying which objects are moving. To do so, we leverage a key assumption that most points in the world are stationary and thus move coherently with a camera movement. Therefore, we can estimate the camera motion between two consecutive frames by estimating the fundamental matrix relating the two. We use a feature detector (*e.g.*, SIFT) to identify keypoints in both frames [4], and use brute force pairwise matching to identify matching keypoints between the two frames. Then, we estimate the fundamental matrix F using a RANSAC procedure that removes outliers, namely pairs of points p_1, p_2 that do not satisfy the equation $p_2^T F p_1 = 0$. Indeed, we find that we are able to obtain a fundamental matrix consistent with most of the keypoints (Fig. 4).

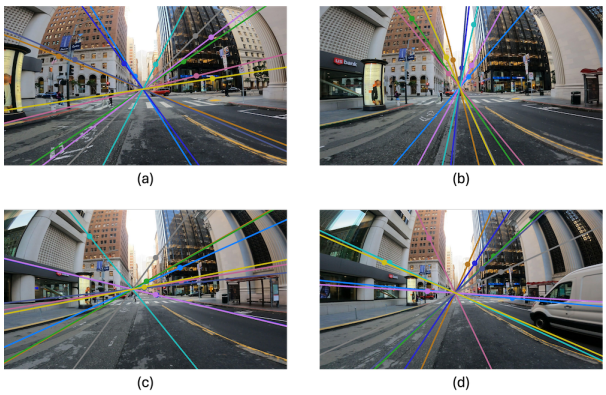

Figure 4. Estimated epipolar lines for sampled keypoints in each frame. Fundamental matrices are estimated between each frame and the preceding frame. Note that consistent keypoints are on static landmarks.

Given an estimated fundamental matrix, we evaluate the consistency of each segmented object with the inferred camera movement. In particular, we calculate the reprojection loss $\frac{\|p_2^T F p_1\|}{\|F p_1\|_2}$ as the distance from p_2 to the corresponding

epipolar line Fp_1. For each segmented object, we classify it as moving if the mean reprojection loss of all keypoints within the segmentation mask exceeds a threshold (Fig. 5).

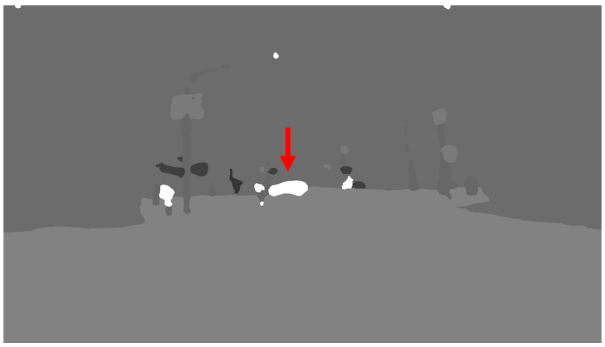

Figure 5. Epipolar reprojection losses for each segmented object in a frame. Brighter corresponds to higher loss. The red arrow points to the segmented car that moves between frames.

2.3. Inpainting

Once moving objects are identified, we can simply remove them with their corresponding segmentation masks, and pass the resultant image to an image inpainting model. Here, we use Stable Diffusion [8], but we note that our pipeline is general to any inpainting model. An example result is shown in Fig. 6, where we see that the moving car and pedestrian have been removed from the scene and inpainted with a plausible natural background. Another example is shown in Fig. 7, where various objects are removed from the scene.

3. Discussion

In this project, we developed a modular pipeline for cleaning StreetView image sequences by combining traditional computer vision ideas such as keypoint detection and fundamental matrix estimation with recent machine learning-based advancements in panoptic segmentation and diffusion inpainting. Our method demonstrates the capability to disentangle static and dynamic components of urban scenes, yielding cleaner imagery suitable for downstream tasks such as virtual navigation and 3D reconstruction.

Our method benefits from both the versatility of modern computer vision models, as well as the robustness of classical computer vision algorithms. Panoptic segmentation models are able to precisely identify objects within the scene and diffusion models are able to generate visually compelling inpaintings from arbitrary masks, enhancing the overall quality of the output frames. At the same time, traditional SIFT keypoint detection, combined with fundamental matrix estimation and epipolar reprojection loss evaluation, enables reliable differentiation between static and moving elements that current machine learning models often struggle at. Together, these methods enable us to perform StreetView cleaning with nearly no constraints as previous methods had: **no camera pose or depth information needed!**

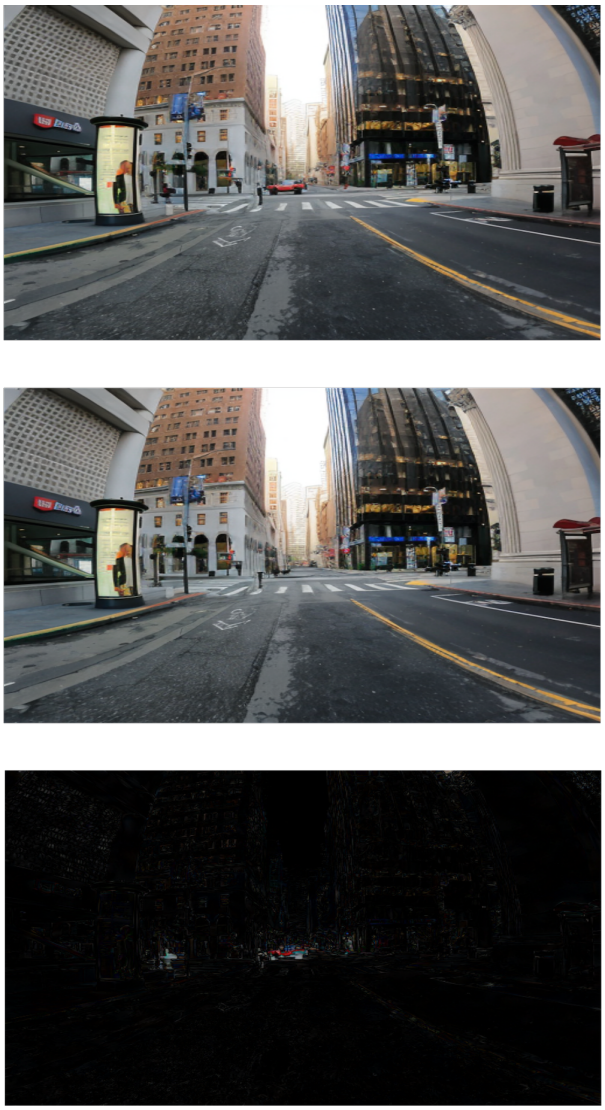

Figure 6. Inpainted StreetView frame. Top: original frame. Middle: inpainted frame. Bottom: difference between the two frames.

Despite these promising results, the pipeline has some limitations. First, the accuracy of moving object identification relies heavily on the robustness of keypoint detection and matching algorithms. Real-world problems such as occlusion, different lighting conditions and camera parameters between frames, or lack of texture may degrade the quality of estimated fundamental matrices and, consequently, the classification of static and moving objects (Fig. 8). Future work could explore the integration of learned feature

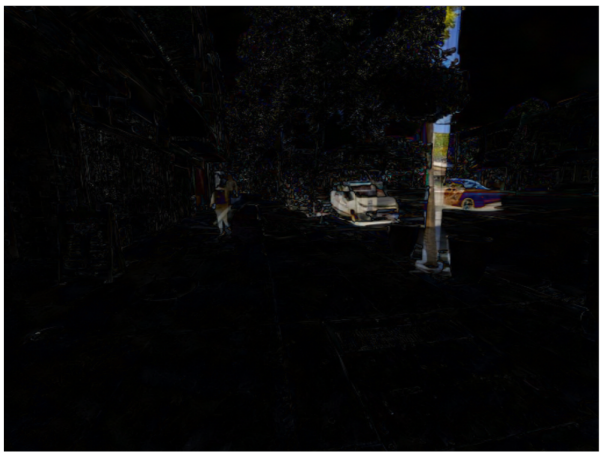

Figure 7. Inpainted StreetView frame. Top: original frame. Middle: inpainted frame. Bottom: difference between the two frames.

descriptors or deep learning-based motion estimation techniques to mitigate these issues.

Figure 8. Example of limitations of our pipeline. Left, inferred object movement mask. Note that the objects within the red bounding boxes are actually static, but were misidentified as moving due to camera variance between the two frames. Right top, original image. Right bottom, inpainted image (failure case).

References

[1] Dragomir Anguelov, Carole Dulong, Daniel Filip, Christian Frueh, Stéphane Lafon, Richard Lyon, Abhijit Ogale, Luc Vincent, and Josh Weaver. Google street view: Capturing the world at street level. *Computer*, 43, 2010. 1

[2] Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. *CoRR*, abs/2112.01527, 2021. 1, 2

[3] Arturo Flores and Serge Belongie. Removing pedestrians from google street view images. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, pages 53–58, 2010. 1

[4] D.G. Lowe. Object recognition from local scale-invariant features. In *Proceedings of the Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157 vol.2, 1999. 2

[5] Ondrej Miksik and Vibhav Vineet. Live reconstruction of large-scale dynamic outdoor worlds. *CoRR*, abs/1903.06708, 2019. 1

[6] Gerhard Neuhold, Tobias Ollmann, Samuel Rota Bulò, and Peter Kontschieder. The mapillary vistas dataset for semantic understanding of street scenes. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 5000–5009, 2017. 1

[7] Angelo Nodari, Marco Vanetti, and Ignazio Gallo. Digital privacy: Replacing pedestrians from google street view images. In *ICPR*, pages 2889–2893, 2012. 1

[8] 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. 1, 3

[9] Ries Uittenbogaard, Clint Sebastian, Julien A. Vijverberg, Bas Boom, Dariu M. Gavrila, and Peter H. N. de With. Privacy protection in street-view panoramas using depth and multi-view imagery. *CoRR*, abs/1903.11532, 2019. 1