The DEVIL is in the Details: A Diagnostic Evaluation Benchmark for Video Inpainting

Ryan Szeto Jason Corso

Electrical Engineering and Computer Science, University of Michigan

Purpose

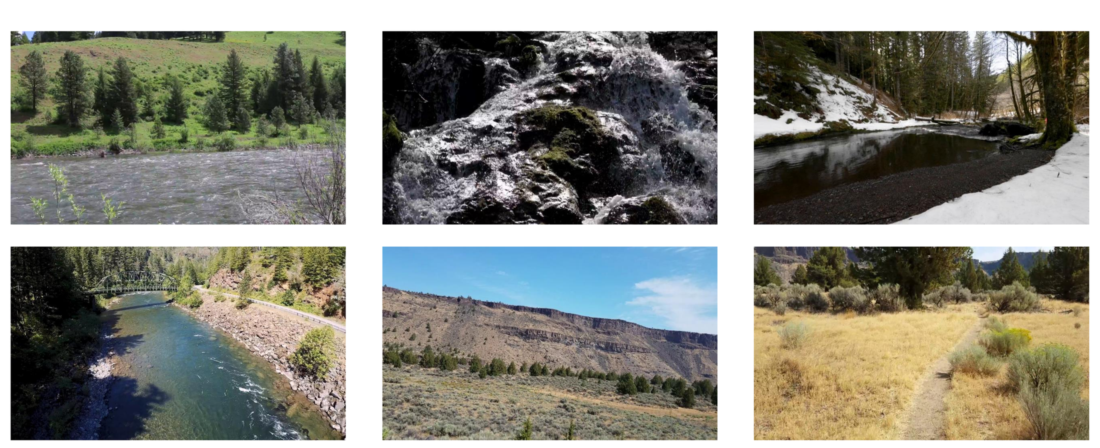
DEVIL is a comprehensive video inpainting benchmark that directly ties video and mask content to inpainting quality.

- In video inpainting, the shape/motion of the missing region and the content/motion of the observed video may affect how shareable appearance information is across frames, and therefore affect inpainting model performance.
- Prior work has not correlated video inpainting model performance with properties of the input video and mask in a quantified, large-scale environment.
- We identify video and mask properties that demonstrably affect inpainting quality, and quantify their impact through a disciplined evaluation scheme applied at scale.
- Our benchmark reveals new insights into modern video inpainting approaches, and serves as a valuable tool for future work.

Source Video Collection

The DEVIL dataset contains 1,250 BG-only scene clips from Flickr.

- Our background-only source videos allow us to study how background content affects video inpainting performance independent of foreground object behavior, and vice-versa.
- To enable scalable background video collection, we sourced videos of natural outdoor scenery from Flickr, which are less likely to contain foreground objects.
- We filtered out foreground objects and shot transitions with automatic methods followed by manual inspection.



Project Page

ryanszeto.com/projects/devil

Acknowledgements

This material is based upon work supported by the National Science Foundation under Grant No. 1628987. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

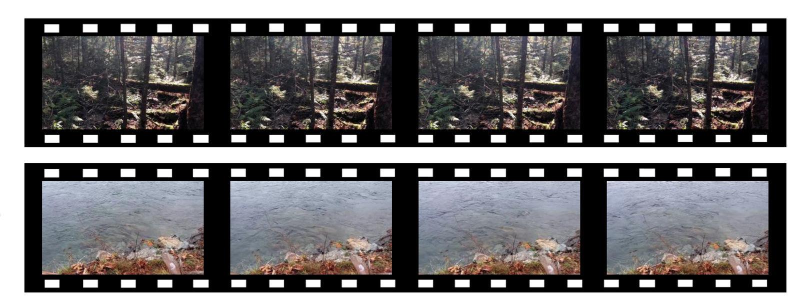
DEVIL Attributes

DEVIL attributes enable disentangled evaluation of FG and BG content on video inpainting quality.

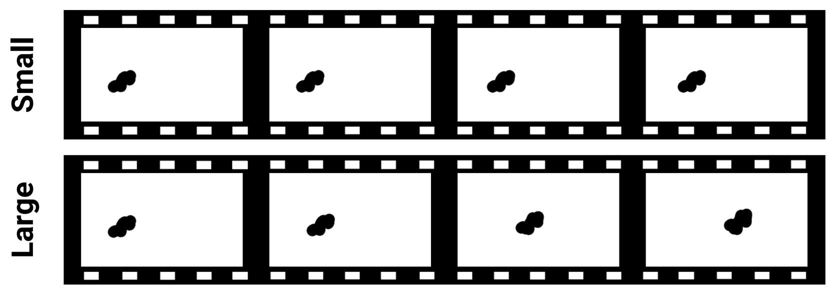
- We distinguish between small and large motion/size since these greatly impact the availability of relevant appearance information in nearby frames.
- We label camera and background (BG) scene motion with a combination of automatic and heuristic approaches. For foreground (FG) masks, we extend the procedural blob generation code from Chang et al. [1] and randomly generate masks with parameters that reflect each attribute setting.

Camera motion
How much the camera pose changes over time

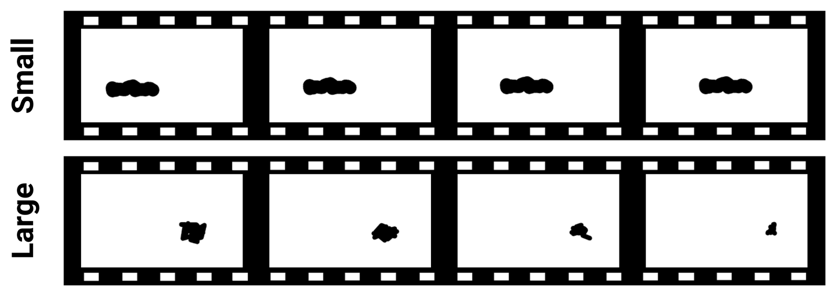
BG scene motion
How much the scene changes independent of camera motion



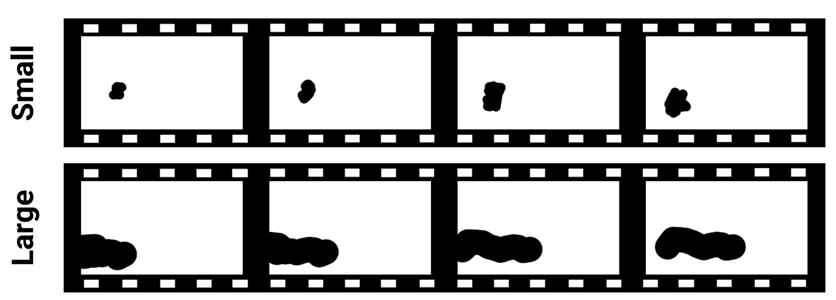
FG displacement
How much the mask's centroid moves relative to the field of view



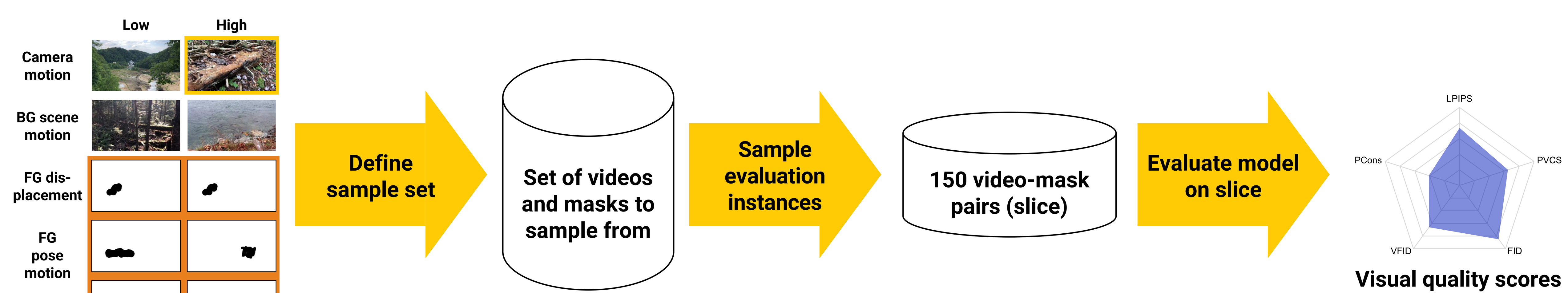
FG pose motion
How much the mask's shape (i.e., outline) changes over time



FG size
How much area the mask takes up in the field of view



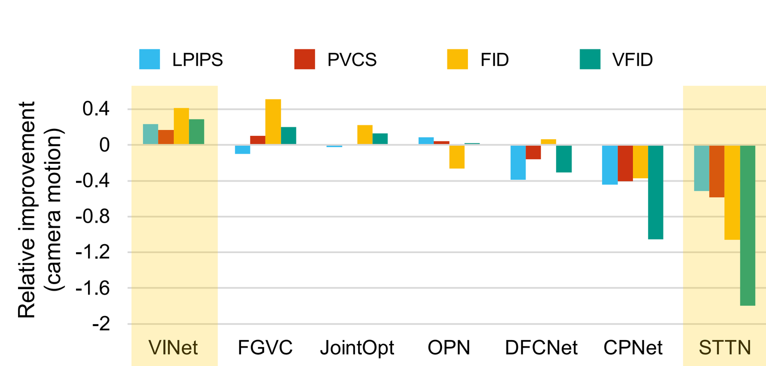
DEVIL Evaluation Scheme



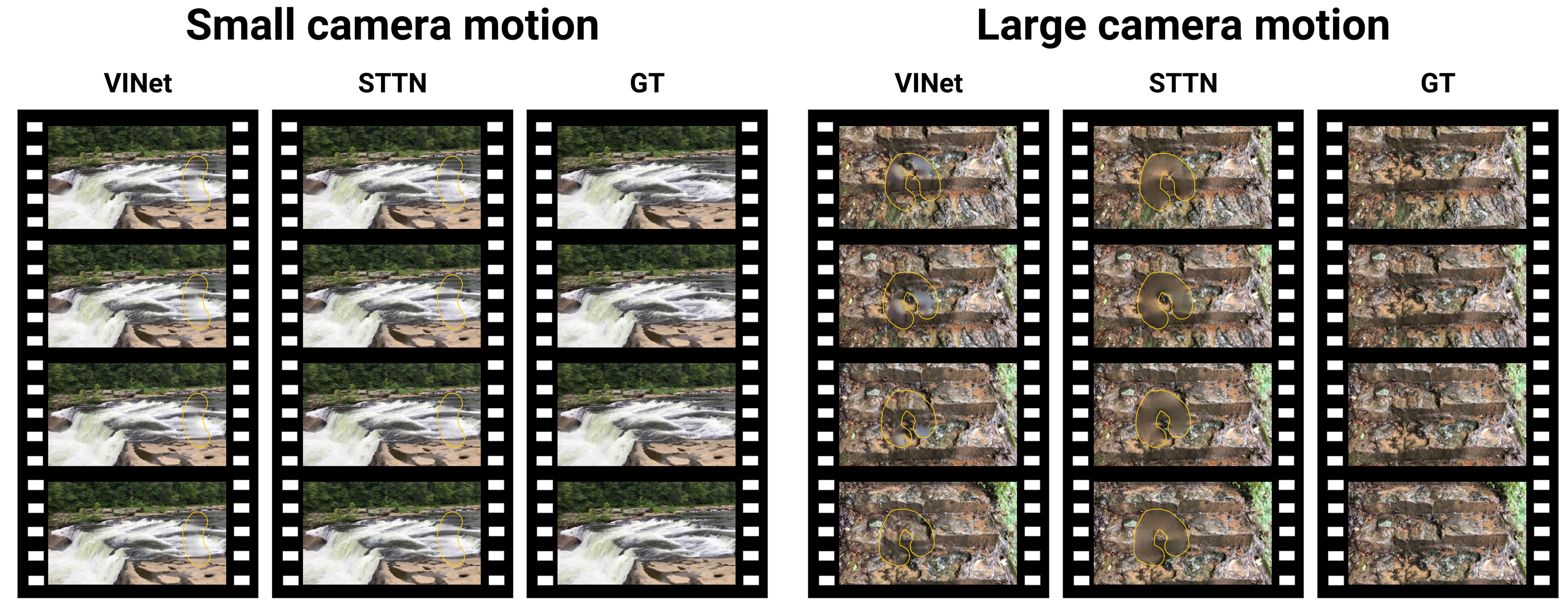
We construct dataset slices, each characterized by one attribute, to see how that attribute affects inpainting performance in isolation. In this illustrative example, we create a slice for high camera motion by sampling all occlusion masks, but only source videos that contain high camera motion. For each model, we randomly sample 150 video-mask pairs per evaluation slice, and then evaluate the model on the resulting set to see how well the corresponding attribute is handled.

Quantitative Failure Case Analysis

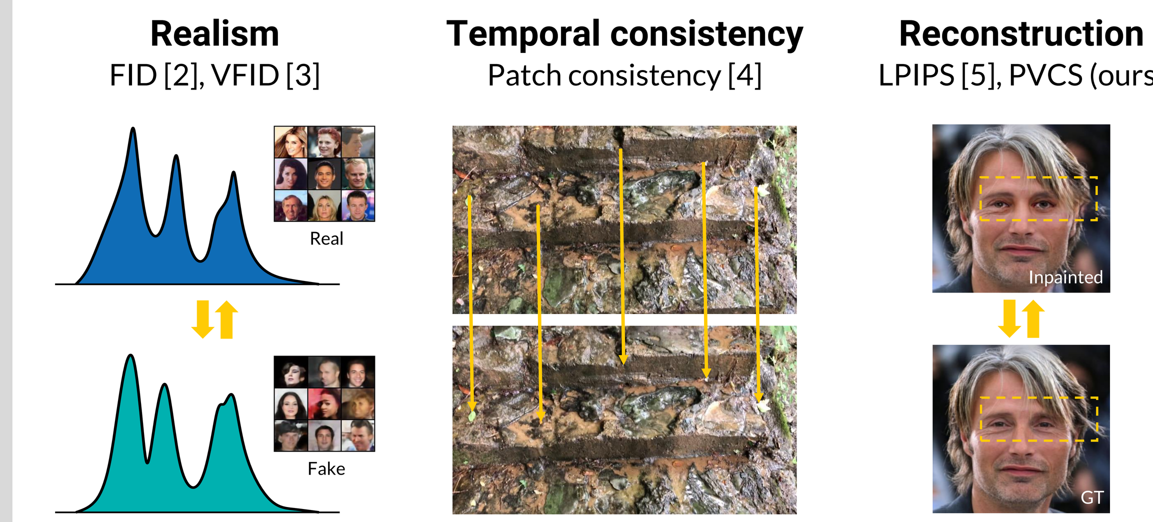
Our quantitative content-based metrics align with qualitative failure cases.



Positive → Better with large camera motion
Negative → Better with camera motion

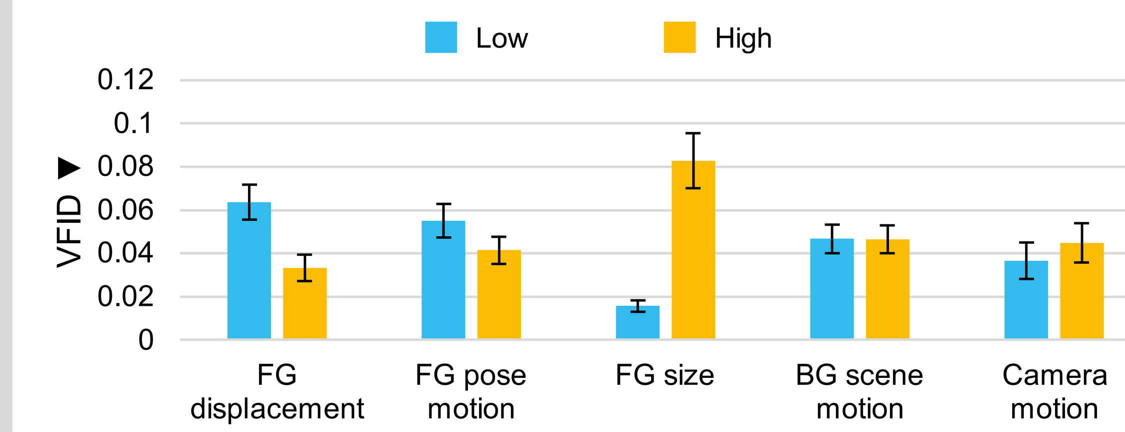


Visual Quality Metrics



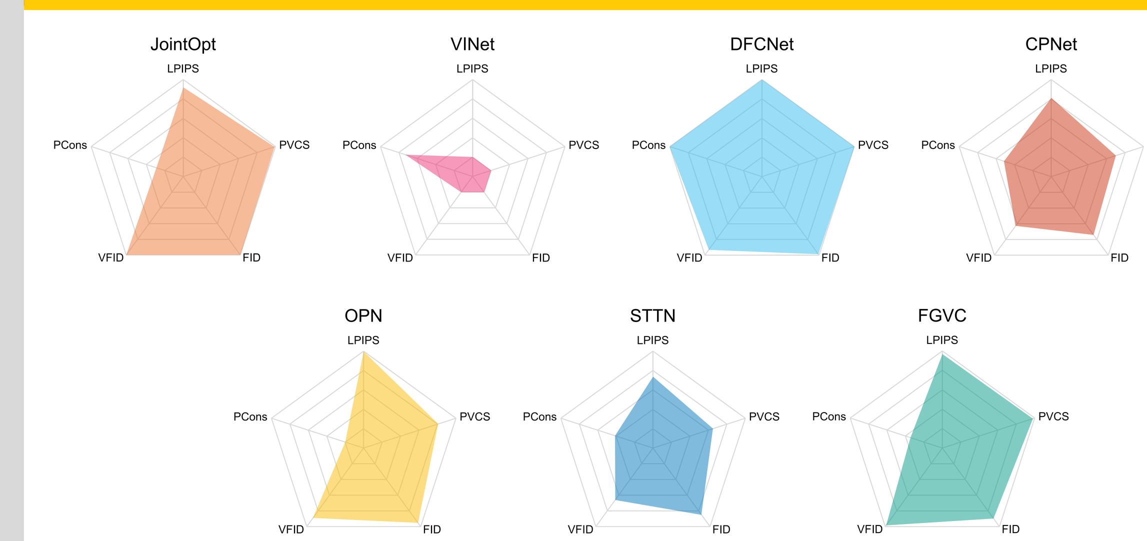
DEVIL Slice Difficulty

Mask properties have a consistent impact on inpainting performance across all models.



Comparison of DEVIL slice difficulty. Lower indicates better inpainting quality.

Radar Plots



Visualizations of each model's performance across the five evaluation metrics averaged over all DEVIL slices; larger area is better. Performance is scaled linearly and independently per metric such that the innermost and outermost pentagons respectively correspond to the weakest and strongest observed mean performance. Models are sorted by publication date.

- **Models that explicitly estimate optical flow produce the best results,** suggesting that flow prediction is key to good video inpainting performance.
- **Non-deep learning approaches perform well,** suggesting that improvements can be made by modernizing older methods instead of just relying on deep learning advances.

References

[1] Chang et al. *Free-Form Video Inpainting With 3D Gated Convolution and Temporal PatchGAN.* ICCV 2019.
[2] Heusel et al. *GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium.* NeurIPS 2017.
[3] Kim et al. *Deep Video Inpainting.* CVPR 2019.
[4] Gupta et al. *Characterizing and Improving Stability in Neural Style Transfer.* ICCV 2017.
[5] Zhang et al. *The Unreasonable Effectiveness of Deep Features as a Perceptual Metric.* CVPR 2018.
[6] Huang et al. *Temporally Coherent Completion of Dynamic Video.* ACM Transactions on Graphics 2016.
[7] Xu et al. *Deep Flow-Guided Video Inpainting.* CVPR 2019.
[8] Lee et al. *Copy-and-Paste Networks for Deep Video Inpainting.* ICCV 2019.
[9] Oh et al. *Onion-Peel Networks for Deep Video Completion.* ICCV 2019.
[10] Zeng et al. *Learning Joint Spatial-Temporal Transformations for Video Inpainting.* ECCV 2020.
[11] Gao et al. *Flow-Edge Guided Video Completion.* ECCV 2020.