Revealing Disocclusions in Temporal View Synthesis through Infilling Vector Prediction

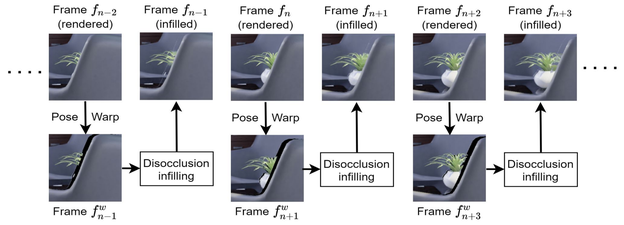
Vijayalakshmi Kanchana **Nagabhushan Somraj** **Suraj Yadwad** **Rajiv Soundararajan**

Indian Institute of Science, Bengaluru, India.

For videos results, dataset, code and more!

Egomotion-Aware Temporal View Synthesis

Predict a future video frame from past frames using depth and egomotion.



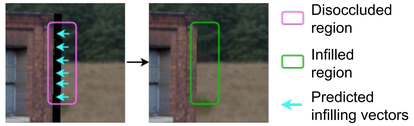
Camera motion disoccludes previously hidden scene content - needs to be infilled.

Contributions:

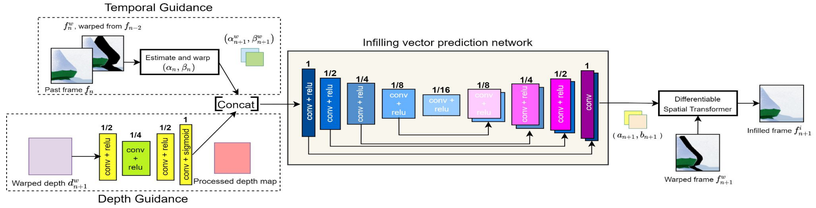
- Infilling vector prediction network for disocclusion infilling with temporal and depth guidance.
- New challenging dataset- IISc-Virtual Environment Exploration Dataset (IISc-VEED).

Infilling vectors:

- Points from known region to disoccluded region in the warped frame.
- Infill disoccluded pixels by copying intensities using the infilling vectors.



Infilling Vector Prediction (IVP)

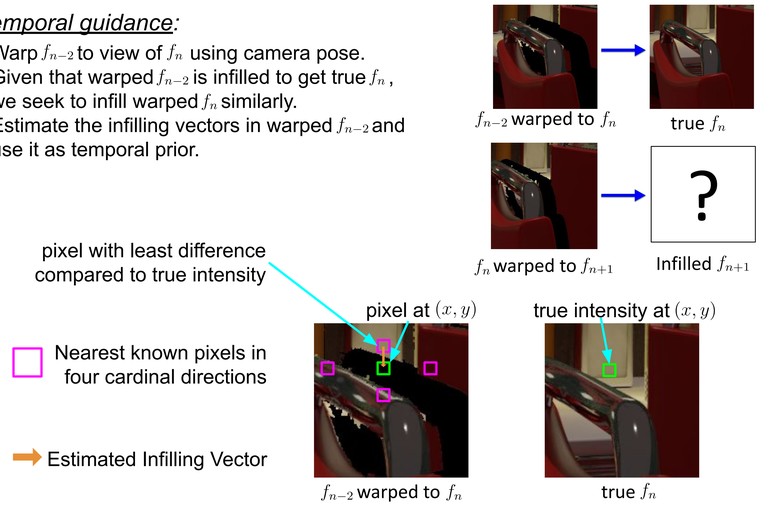


- Given frame f_n, depth d_n, and the relative camera transformation T from f_n to f_{n+1}, we use projective geometry based warping to warp f_n to f_{n+1}- this creates disocclusions.
- We input the temporal and depth priors to a U-Net and predict infilling vectors.
- Training loss: $\mathcal{L} = \lambda_1 \mathcal{L}_{\text{MSE}}(f^i_{n+1}, f_{n+1}) + \lambda_2 \mathcal{L}_{\text{SSIM}}(f^i_{n+1}, f_{n+1}) + \lambda_3 \mathcal{L}_{\text{smooth}}(a_{n+1}, b_{n+1})$
  where mean squared error ($\mathcal{L}_{\text{MSE}}$) and structural similarity ($\mathcal{L}_{\text{SSIM}}$) are computed between the infilled and true frames and $\mathcal{L}_{\text{smooth}}$ is the edge-aware smoothness loss on predicted infilling vectors.

Temporal and Depth Guidance

Temporal guidance:

- Warp f_{n-2} to view of f_n using camera pose.
- Given that warped f_{n-2} is infilled to get true f_n, we seek to infill warped f_n similarly.
- Estimate the infilling vectors in warped f_{n-2} and use it as temporal prior.



Depth Guidance:

- Disocclusions typically belong to background regions.
- Provide normalized depth map as input.
- This guides the network to copy intensities from the relative backgrounds.

IISc-VEED: New Dataset

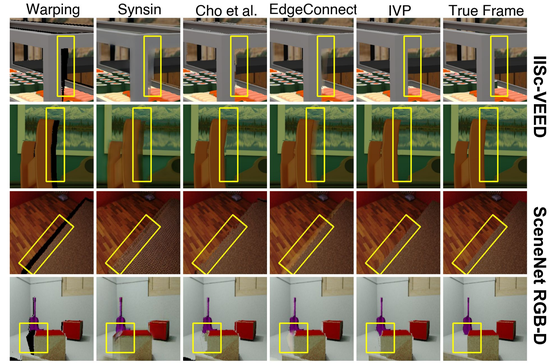
- 800 videos, 1920x1080, 30fps, 12 frames per video.
- Videos of indoor and outdoor scenes, rendered with Blender.
- Camera trajectories chosen to be realistic and produce challenging disocclusions.
- Provide RGB frames, depth, camera extrinsics and intrinsics.

RESULTS

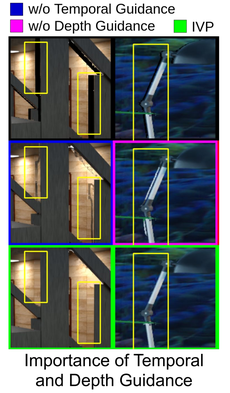
Benchmarks:

- Novel View Synthesis: **SynSin** [1]
- Depth Image-Based Rendering: **Cho *et al.*** [2] (Input is warped frame)
- Image Inpainting: **EdgeConnect** [3] (Input is warped frame)

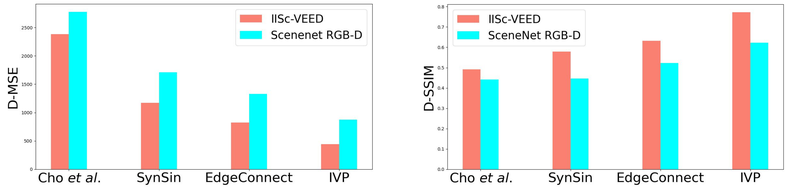
Datasets: Our IISc-VEED dataset and the SceneNet RGB-D [4]



Qualitative comparisons on IISc-VEED and SceneNet RGB-D



Importance of Temporal and Depth Guidance



Quantitative comparisons on IISc-VEED and SceneNet RGB-D

REFERENCES

[1] Wiles *et al.* "Synsin: End-to-end view synthesis from a single image". CVPR 2020.
[2] Cho *et al.* "Hole filling method for depth image based rendering based on boundary decision". IEEE Signal Processing Letters, 2017.
[3] Nazeri *et al.* "Edgeconnect: Structure guided image inpainting using edge prediction". ICCVW 2019.
[4] McCormac *et al.* "Scenenet RGB-D: Can 5m synthetic images beat generic imagenet pre-training on indoor segmentation? ". ICCV 2017.

Acknowledgements: This work was supported by a grant from Qualcomm.