Live Video Enhancements using Monocular Depth Estimation

Progress Updates

1. Literature Review

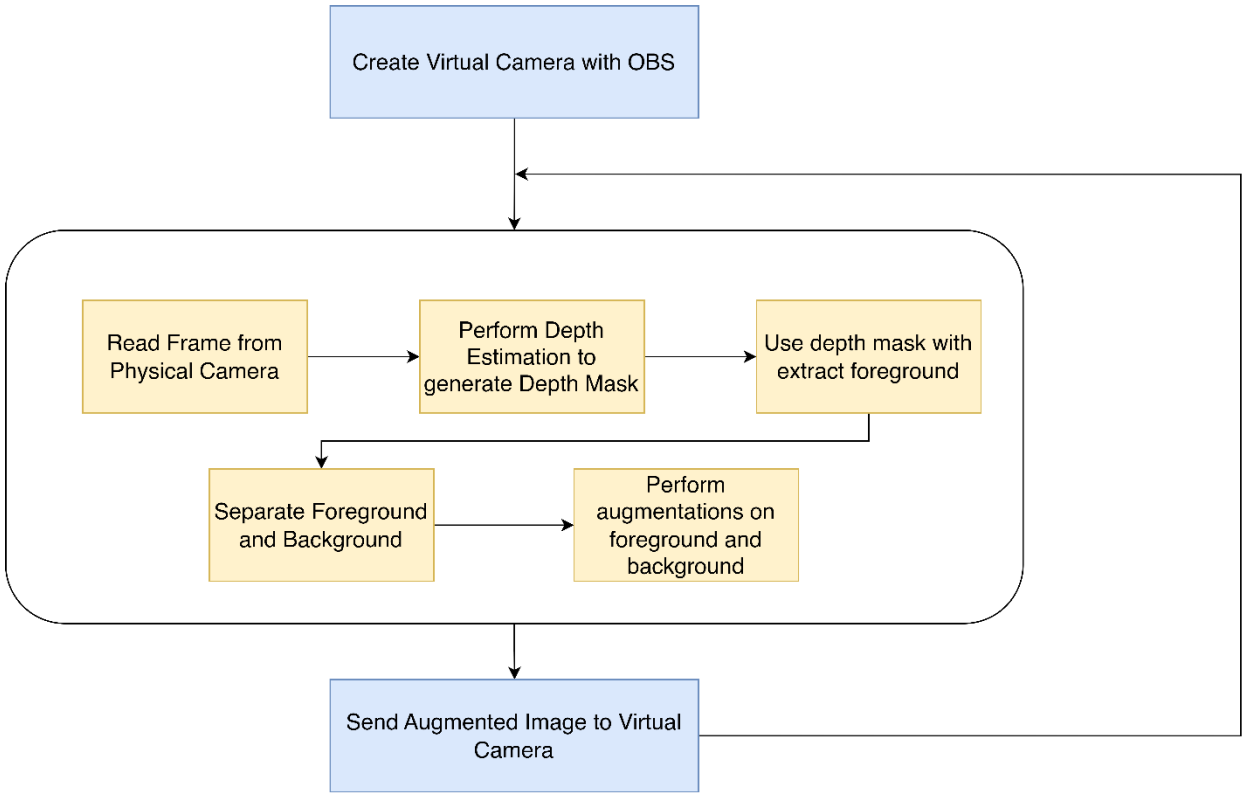
Depth estimation with depth-centric cameras is not a new technology, it had reached a level of maturity where it was being used for casual gaming with Microsoft XBOX 360 Kinect [1]. However, the usage of RGB imagery to attain depth masks is one which is catching traction as of now. Previously, pairs of RGB images (stereoscopic imagery) have been used to generate a depth map by employing the principles of binocular disparity [2]. Current works eliminate the usage of the secondary camera by utilizing a monocular approach, thus giving rise to Monocular Depth Estimation.

Depth estimation on its own has many uses in robotic navigation, autonomous driving, and virtual reality [3]. Such uses cases often utilize application based sensors such as LIDAR sensors that are being used in autonomous vehicles [4,5]. Such sensors are usually not cheap and require custom pipelines to generate ground truth with annotation in the data. This often makes the collection of accurate large & varied ground truths a challenging task, not to mention, requiring significant monetary efforts.

While there may be works with depth estimation in domains related to autonomous vehicles, there exists few prior works that focus on using depth estimation for real-time video enhancements. This again ties in with the fact that there does not exist a dataset to support this task. The works of Godard et al. [4] focus on a self-supervised methodology for Monocular Depth Estimation, which aims to reduce the effort for ground truth generation.

On the contrary from depth estimation, a lot of work has been done to enhance and improve the video quality in real time. The works of Yeo et al. [6] propose a scalable neural network-based enhancement model that is able to restore video quality that is deteriorated by reduced bandwidth. They also highlight a fact that around 50% of stream viewers tend to abandon a live stream when there is quality degradation for more than 90 seconds. The works of Valanarasu et al. [7] focus on video restoration by using a dual approach for restoration where they use a convolutional encoder for feature extraction and MLP based mixing blocks. However, they base their benchmarks on the NVIDIA A100 GPU, which is usually used in datacenters for research and is not available or feasible to be used by the average consumer.

2. Proposed Pipeline



3. Potential Hurdles

- Running a depth estimation model on each frame would yield performance that is not suitable for live video augmentation.
- There exists no dataset geared towards near-field depth maps, existing datasets have imagery from the streets and the indoor objects (KITTI dataset and NYU Depth v2 respectively). There might be issues adapting these datasets to human torsos.
- Packaging the entire codebase into a single executable may not be possible due to GPU dependencies.
- There is no direct prior work for metric comparison, will have to rely on computational requirements to generate a comparison from previous models.

4. Timeline Adherence

Timeline		Task Description	Status
March	March 03	**Project Proposal**	✓

	March 20	Review model architectures and finalize the viability of the project	✓
April	April 09	**Project Progress Report**	✓
	April 14	Finalize Depth Estimation based Backend	
	April 15	Start work on the pipeline that feeds into the backend	
	April 28	Polish the demo based on combined frontend and backend	
May	May 05	**Project Final Presentation**	

5. References

1. Azure Kinect DK – Develop AI Models | Microsoft Azure. (n.d.). https://azure.microsoft.com/en-us/products/kinect-dk/
2. Aslam, A., & Ansari, M. (2019). Depth-map generation using pixel matching in stereoscopic pair of images. arXiv preprint arXiv:1902.03471.
3. Ming, Y., Meng, X., Fan, C., & Yu, H. (2021). Deep learning for monocular depth estimation: A review. Neurocomputing, 438, 14-33.
4. Godard, C., Mac Aodha, O., Firman, M., & Brostow, G. J. (2019). Digging into self-supervised monocular depth estimation. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 3828-3838).
5. Geiger, A., Lenz, P., Stiller, C., & Urtasun, R. (2013). Vision meets robotics: The kitti dataset. The International Journal of Robotics Research, 32(11), 1231-1237.
6. Yeo, H., Lim, H., Kim, J., Jung, Y., Ye, J., & Han, D. (2022, August). NeuroScaler: neural video enhancement at scale. In Proceedings of the ACM SIGCOMM 2022 Conference (pp. 795-811).
7. Valanarasu, J. M. J., Garg, R., Toor, A., Tong, X., Xi, W., Lugmayr, A., ... & Menini, A. (2023). ReBotNet: Fast Real-time Video Enhancement. arXiv preprint arXiv:2303.13504.