**Live Video Enhancements using Monocular Depth Estimation**

1. **Problem Description**

In today’s world, there has been a lot of work done toward facial localization and tracking of faces. As a matter of fact, this technology was available in a limited capacity in the days of digital camcorders (the late 2010s). With our rising hardware-enabled capabilities, object tracking is a problem, that can be considered “solved”. Tech giants such as Apple make use of depth & infrared cameras to localize & generate a facial mapping of our faces and use that for tracking and authentication [1]. Such sensor suites are also used in a feature on Apple products, known as Center Stage, where a user is tracked in real-time using their “ultra-wide true-depth camera” [2] with the use of machine learning. Such features are made possible by the use of equally rivaling hardware.

However, when a supply chain is disrupted, as it was during the COVID-19 pandemic, the disruption in the availability of silicon chips led to a direct shortage of electronic components. This affected multiple sectors of the economy and essentially enforced a pause in multiple industries. This highlights a key flaw in the business, where a feature is made possible entirely by the use of certain hardware components.

This work attempts on alleviating the use of depth-centric cameras by the creation of a system that uses a standard RGB image camera to generate a depth map. This depth map can be utilized to isolate the focused subject from the background and perform a limited tracking operation on the subject as well.

1. **Related Works**

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 [3]. 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 [4]. Current works eliminate the usage of the secondary camera by utilizing a monocular approach, thus giving rise to Monocular Depth Estimation.

1. **Datasets**

* **NYU Depth v2**

It is a collection of indoor RGB images, with corresponding depth and segmentation ground truths taken at a spatial resolution of 640x480 pixels [5]. It is comprised of imagery from the Microsoft Kinect, and these were taken in multiple residential and commercial locations.

* **KITTI**

It is a dataset of imagery taken in different outdoor scenes with the use of a LIDAR sensor [6]. The raw datasets feature imagery of 1241x376 pixels and scenes of different categories such as ‘Road’, ‘City’, ‘Residential’, ‘Campus’, and ‘Person’.

1. **Timeline**

| **Timeline** | | **Task Description** |
| --- | --- | --- |
| March | March 3 | **Project Proposal** |
| March 20 | Review model architectures and finalize the viability of the project |
| March 31 | **Project Progress Report** |
| April | 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 5 | **Project Final Presentation** |

1. **Final Deliverables**

Since most teleconferencing software have customizability that allows the user to choose their input devices, the final expectation from this work is an end-to-end deployment that can be used by any end user as an augmentation for their video feed to a virtual webcam. This work would be available publicly on GitHub and made available on DockerHub for demonstration purposes.

1. **References**
2. Apple. (2022, April 27). *About Face ID advanced technology*. Apple Support. <https://support.apple.com/en-us/HT208108>.
3. Adorno, J. (2021, May 19). Roundup: Here’s how the 2021 iPad Pro Center Stage feature really works. 9to5Mac. <https://9to5mac.com/2021/05/19/roundup-heres-how-the-2021-ipad-pro-center-stage-feature-really-works/>
4. Azure Kinect DK – Develop AI Models | Microsoft Azure. (n.d.). <https://azure.microsoft.com/en-us/products/kinect-dk/>
5. Aslam, A., & Ansari, M. (2019). Depth-map generation using pixel matching in stereoscopic pair of images. arXiv preprint arXiv:1902.03471.
6. Silberman, N., Hoiem, D., Kohli, P., & Fergus, R. (2012). Indoor segmentation and support inference from rgbd images. ECCV (5), 7576, 746-760.
7. 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.