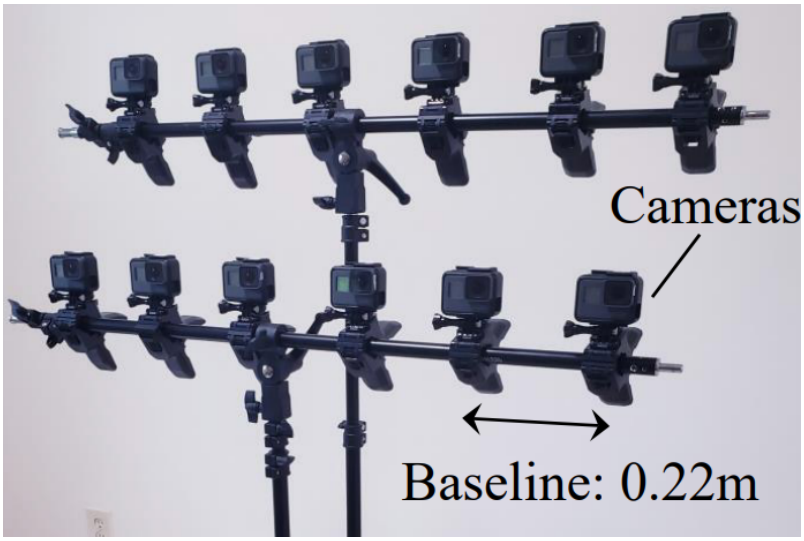
Novel View Synthesis of Dynamic Scenes with Globally Coherent Depths from a Monocular Camera[1]

Mohammadreza Asherloo
Department of Mechanical, Materials and Aerospace Engineering
Illinois Institute of Technology

December 2, 2020

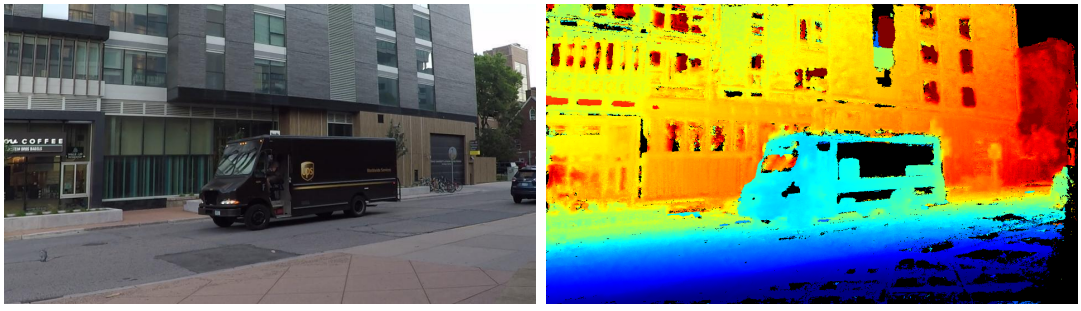
Concluding Report:

Synthesising dynamic scenes using 12 images from a dynamic scene with 12 different cameras. The difficult part is estimating the depth of dynamic scene from static cameras because objects in the scene are moving so the dense or sparse reconstruction will not work good enough when used alone.
To address this issue, dynamic scene was captured using 12 cameras aligned in two rows as follows:

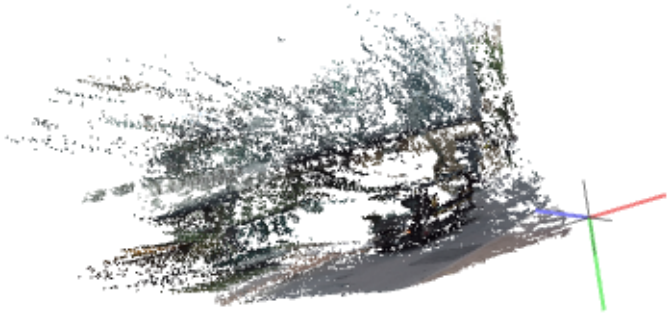


By using these cameras, 12 images were taken from 12 different static scenes in a short time to make a dynamic scene. From these 144 images, 12 images were selected as the representative of the dynamic scene and the input to neural network:

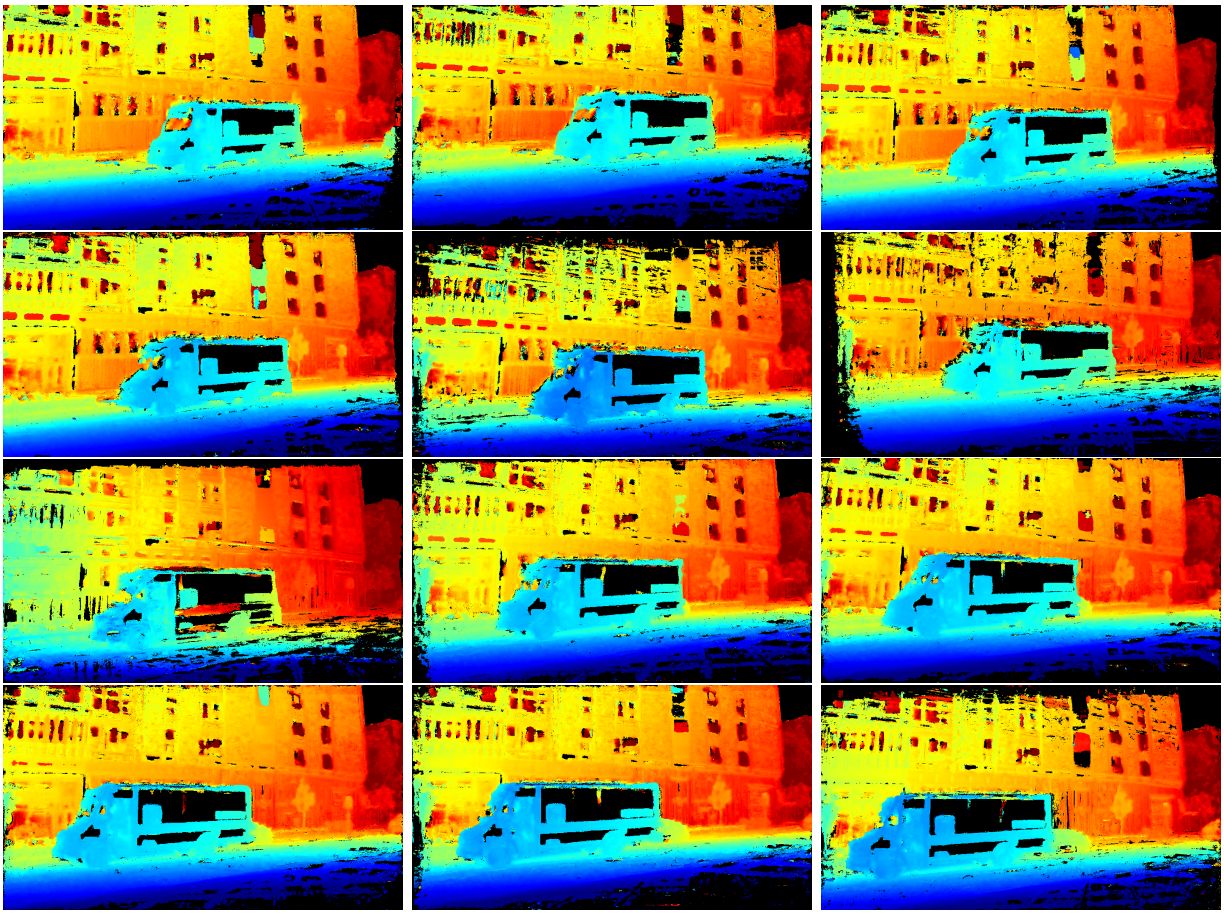
12 images from each static scene was used to reconstruct the scene and estimate the depth of scene. These depths will be used later to have a consistent background and divide the foregrounds from backgrounds in images and make a dynamic scene.
An example of 3D reconstructed scene and estimated depth visualized on the corresponding image is as follows:



This 3D reconstruction using 12 images was done using COLMAP software[2]. COLMAP is a general-purpose Structure-from-Motion (SfM) and Multi-View Stereo (MVS) pipeline with a graphical and command-line interface. It offers a wide range of features for reconstruction of ordered and unordered image collections.
12 images from each where fed to COLMAP software and a depth estimation for each scene was produced along with the dense and sparse reconstruction of each scene:

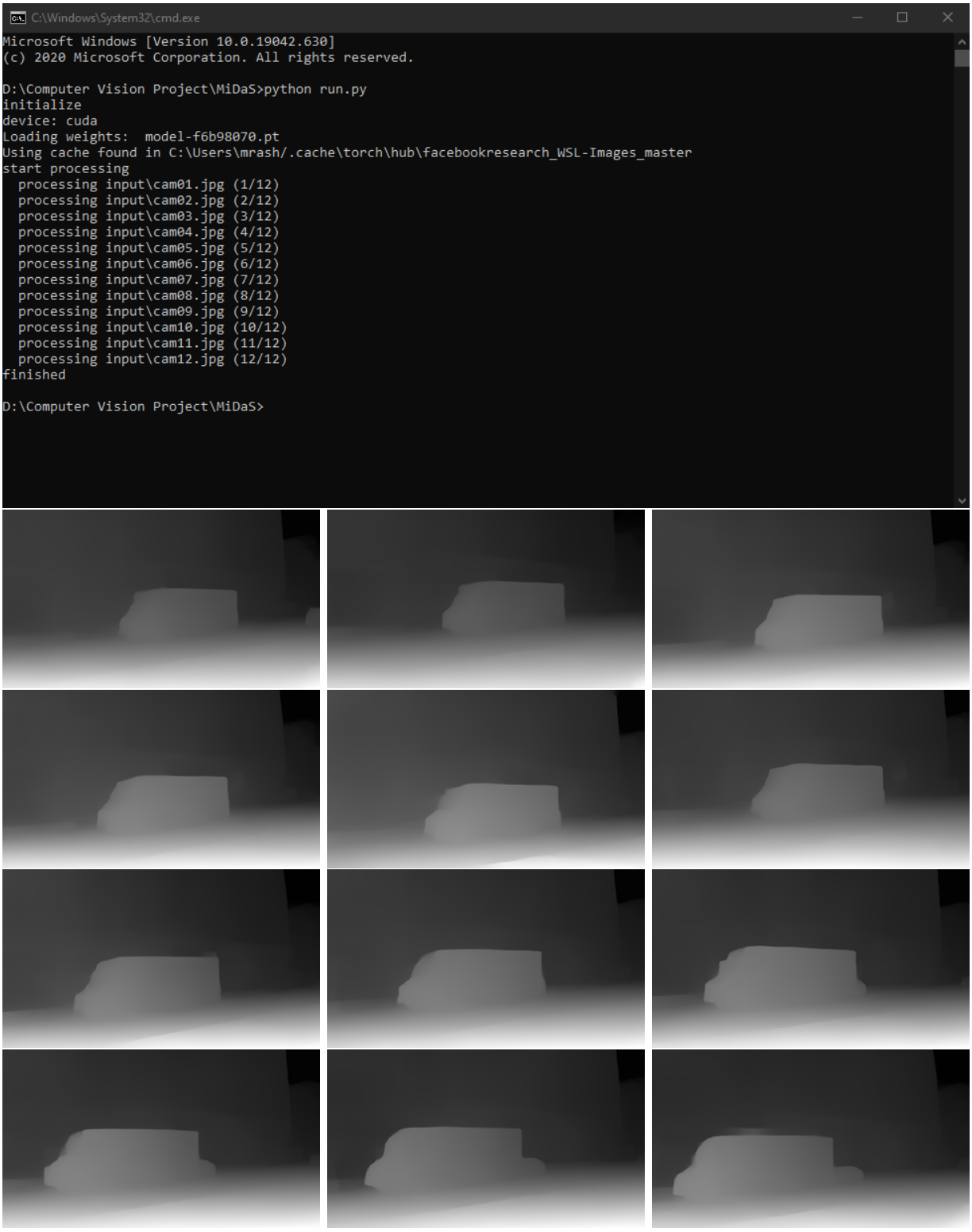


Now 12 input images and 12 depth estimated images are in possession:



Moreover, a depth estimated image from static view is needed. These images were produced using MiDaS project[3]. MiDaS v2.1 was trained on 10 datasets (ReDWeb, DIML, Movies, MegaDepth, WSVD, TartanAir, HRWSI, ApolloScape, BlendedMVS, IRS) with multi-objective optimization. Using MiDaS is fairly simple. The input images should be placed in the input folder and the program should be run through command line and ouput

images will be created and placed in output folder:

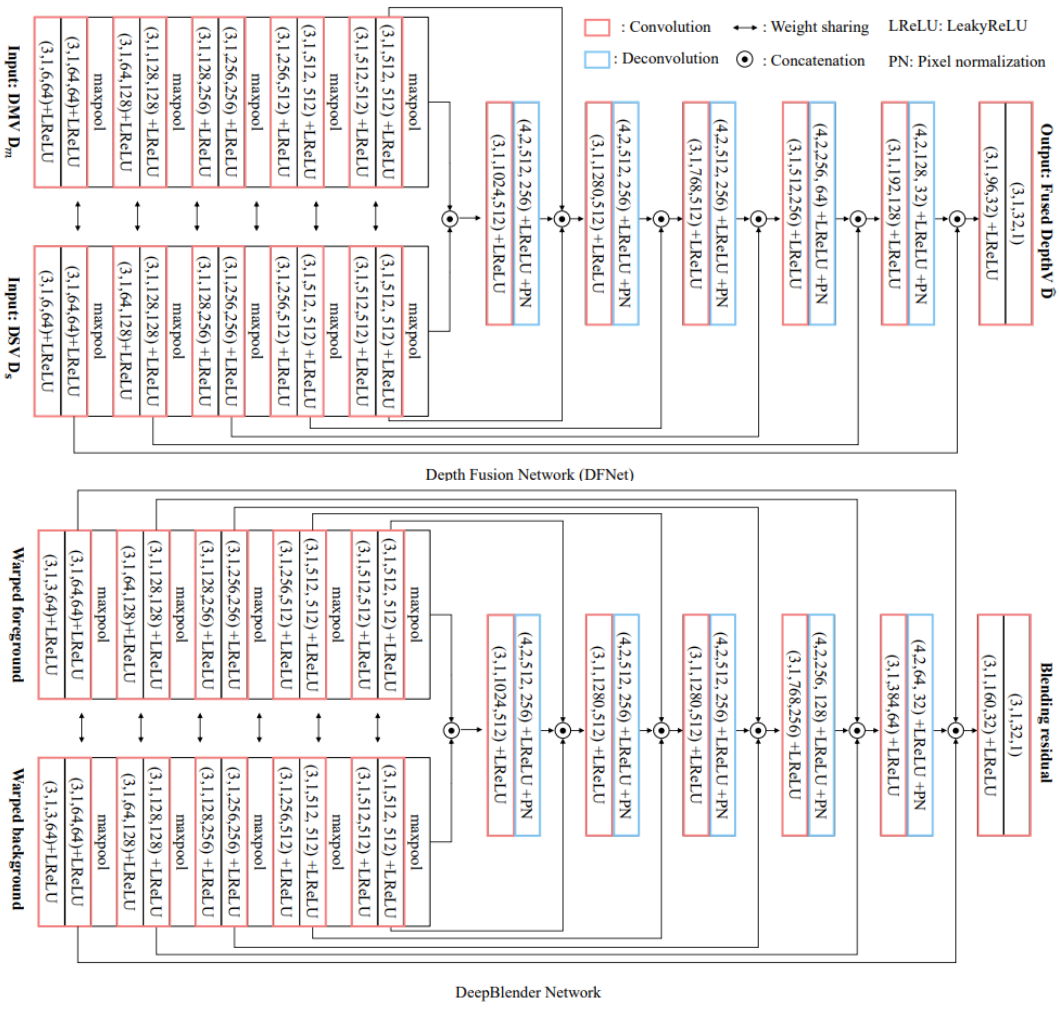


For training the neural net, we will need ground truth depth estimations and masked foreground and background to calculate the loss. For ground truth images, gray-scale depth

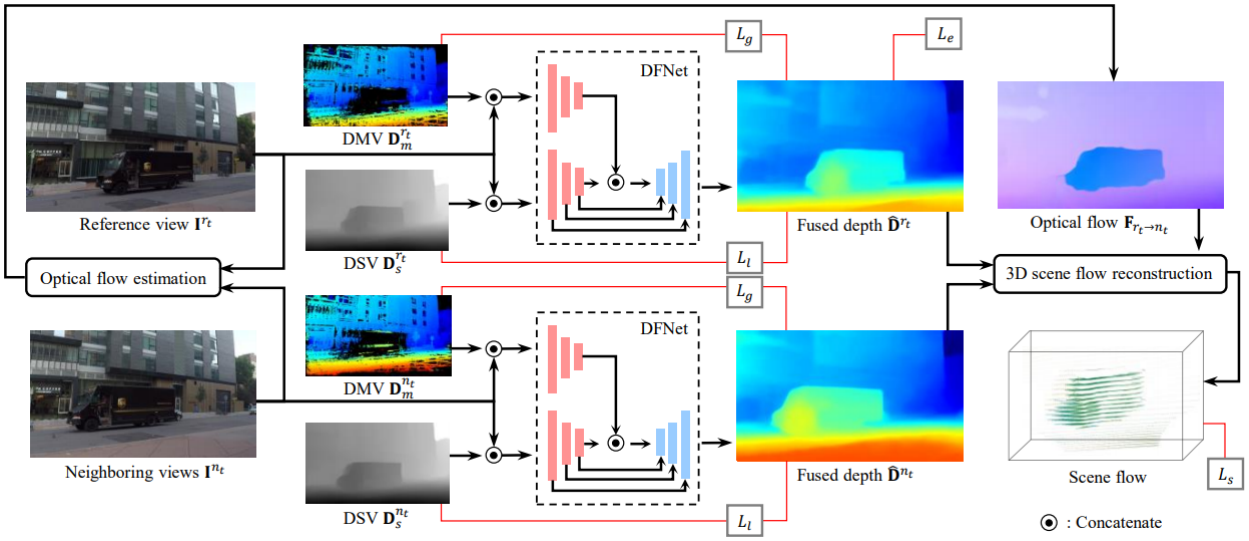
estimated images were used and for mask images, binary (background-foreground) images were used for each scene such as:

After preparing input data, we constructed the pipeline for producing the depth estimated images. The pipeline was working fine but due to lack of hardware capacity, we couldn't train the models. We used google COLAB to train this neural network which crashed because of lack of RAM.

First model has roughly 38 million parameters and the second model has roughly 41 million parameters to train. The model graph for depth fusion (left image) and image blender (right image) based on model architecture provided by authors:

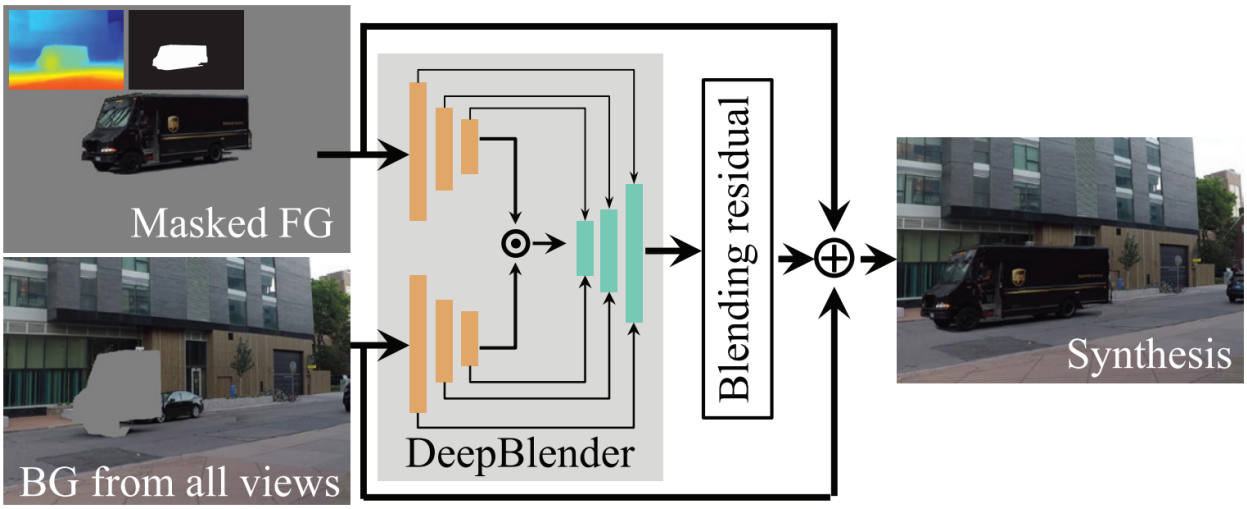


By using keras functional API, we reconstructed the pipeline using concatenation of layers and shared weights. The model architecture is provided in the data folder by names of "Blender.png" and "Depth Fusion.png".
Then these neural networks were used in the depth estimation pipeline as follows:



Then we used foreground masked images:

and the overall background to produce the final images:



Link to dataset: Dataset-clickable.
Link to model weights for MiDaS network: Model weights-clickable.

References

[1] Jae Shin Yoon, Kihwan Kim, Orazio Gallo, Hyun Soo Park, and Jan Kautz. Novel view synthesis of dynamic scenes with globally coherent depths from a monocular camera. June 2020.

[2] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[3] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2020.