CSE353 Milestone 3: Progress Report

Hyeon Joon Lee

DongHun Kim

Scope and Purpose

The primary objective of our group's final project is to enhance the performance of existing approaches of frame interpolation. Many notable approaches rely on 3D images and depth data that can be obtained from them. Some others, who employ 2D images, also estimate the depth and use the data for obtaining the loss function. Having initially planned to incorporate such depth data from 3D videos into our algorithm, our group has shifted to using human join data.

The remaining sections of this paper will explain the progress up until Milestone 3, some successes, in addition to failures. Furthermore, we will state what we aim to achieve by Milestone 4 and our plans.

Progress

Up until now we have come up with an algorithm to first divide a video file into each frame, second detect human existence in that frame, and finally prepare it to be used as a viable training, validating, or testing dataset, upon human detection. We first inspected each and every frame of a video file obtained from YouTube and applied YOLO v3 to check if there is at least one human figure in it. If so, we then apply Openpose to effectively retrieve information about body keypoints, or joints, on each skeleton detected. Retrieving the data as an ndarray, we saved such data for each frame in a separate json file, for later use.

The above algorithm can be found in the file *create_dataset.py.*

Next, we have modified the training algorithm used by SuperSloMo by adding an extra loss function regarding the joint data obtained. We have compared such data from both the ground-truth intermediate frame and the interpolated frame, and added the mean square error as a component of the global loss function, also composed of other four loss functions, namely Reconstruction Loss, Perceptual Loss, Warping Loss, and Smoothness Loss. This can be observed in the file *train.py.*

Success & Failure

As stated in the Progress section, we have managed to retrieve both the input image data and the joint data for training, validating, and testing the model, from raw videos. We believed that saving the join data for each frame in a separate json file would be the optimal option, as we could access it at any time. Nevertheless, the algorithm by SuperSloMo, which we refer to, does not incorporate consecutive frames when choosing the intermediate target and such choosing is completely at random. For instance, while the first frame is indexed 0 and the last frame 8, the intermediate frame can take any index from 1 to 7. Even more, such index is then updated, making it more difficult to figure out the exact frame. Thus, we concluded that it would be more efficient to generate joint data after choosing the frame in the training phase.

Neither the gradient of descent in the loss function nor the psnr (peak signal-to-noise ratio) that we use to evaluate the performance of the model was satisfactory. While this could have resulted from the introduction of our loss function, we believe that the most crucial factor is that we only conducted limited epochs on limited data. More importantly, filtering out only the frames with human detection resulted in a very serious discontinuity between frames. For

instance, if there is no human for 10 seconds in a 30fps video, then there is a loss of 300 frames. When such distinct frames are chosen as the surrounding frames for our interpolated frame, it is obvious that even our ground truth frames will be of no use at all. Indeed, even without the additional loss function, the original algorithm hardly performed any better.

Figure 1 and 2 each represents the loss function and psnr for training without and with keypoint values and their corresponding loss function. There is a huge inconsistency in the loss function for training with keypoint values. We believe this is mainly due to highly distinctive frames chosen as surrounding frames for the interpolated frame. Also, please note that keypoint values and their loss function was only used in the training phase, not in the validating period, due to unexpected increase in the required memory in the GPU. This might have affected the outcome too; we will find out.

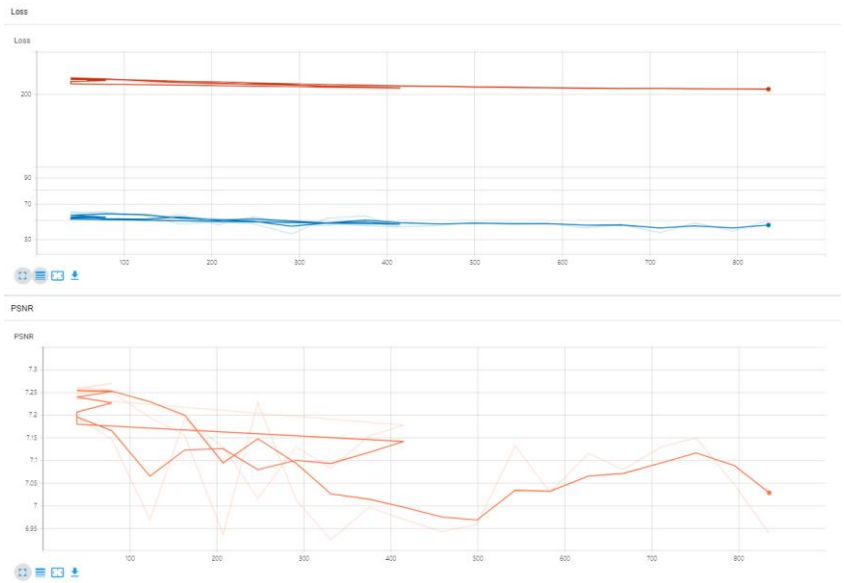

Fig 1: Without Keypoint Values

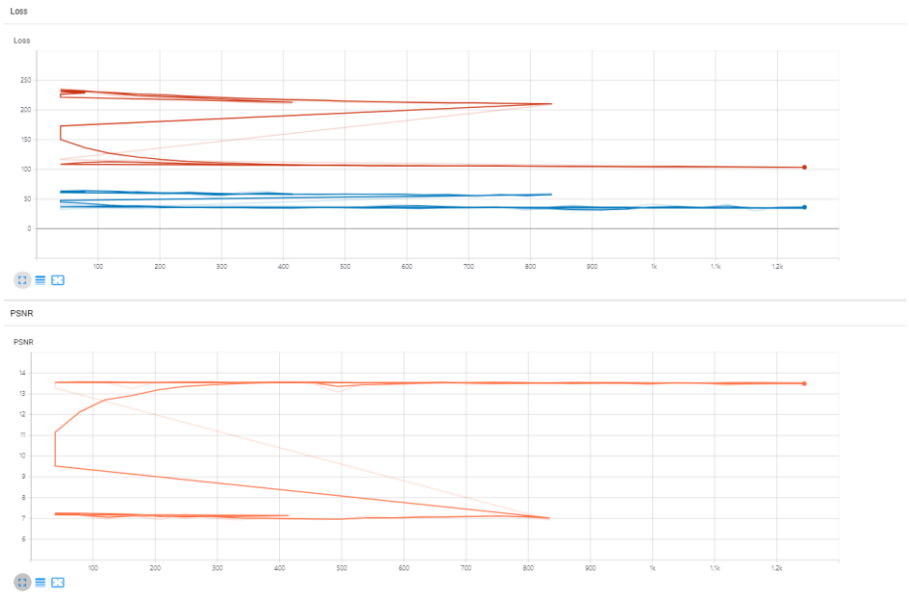

Fig 2: With Keypoint Values

Plan

We will dig more into the underlying concept in our model. Furthermore, we will identify what led to unsatisfactory decline in the loss function and incline in the psnr. We will identify all issues, including the unadjusted coefficients in the new global loss function and distinct frames used in interpolation, and fix it by Milestone 4.