Deep Monocular Depth Estimation for Movies

Owen Thompson

Rochester Institute of Technology
Advisors: Ricardo Figueroa, Flip Phillips, Guoyu Lu

Abstract

Motion picture depth map datasets are not common. This makes application of depth prediction to motion pictures difficult. An indirect unsupervised model has been devised to get around this limitation and a new 3D movie dataset has been assembled for the purpose of training it. Our dataset consists of stereo frames extracted from 3D motion pictures spanning multiple genres and containing diverse content. The proposed model uses an established loss and network architecture which accounts for motion picture stereo disparity and utilizes a monocular image and optic flow as input. This has created a depth estimating tool suitable for applications such as automatic 3D stereo synthesis from 2D input. To the best of our knowledge, the techniques employed by this model have not been leveraged in the way proposed herein and its potential creative utility has not fully been realized elsewhere.

`https://github.com/oxt3479/3D_learner`

Keywords— Deep Learning, Depth Estimation, Motion Pictures, Stereo 3D, Optic Flow, Virtual Reality

1. Introduction

Depth datasets already exist. However, many of them consist of sparse or incomplete maps which are insufficient for training a one-to-one image depth prediction model. The best amongst these existing datasets additionally fail to represent the creative scene content found in motion pictures. [3] Creative scenes call for a network which uses a highly diverse dataset trained using a semi-unsupervised approach. Semi-unsupervised models have been used in the past, but have not been applied to motion pictures. [4, 1]

1.1. Creative Scenes and Complications

In this paper, a creative scene refers to a typical image composition found in a motion picture or film. Common aspects of these frames include: human subjects, nuanced lighting configurations, extensive CGI, and highly variable camera motion and parameters. All of these attributes contribute to an increasingly complicated means of depth prediction. Combating this level of variability calls for a large dataset. Our collection of 1.7 million stereo frame pairs seeks to satisfy this issue. 3D stereo pairs can be extracted from any 3D movie, and each film can provide anywhere from 100,000 to 250,000 frame pair samples. Our dataset currently contains 12 movies.

1.2. Applications and Demonstration

A depth map can be used creatively in many different ways. It permits the changing of image focus. It can also be used when compositing two shots together, allowing subjects to pass in between each other, unlike a simple green-screen. It can be used in applying effects that need depth information to render properly, for instance fog which gets more opaque at further distances. Depth maps can be estimated without automatic monocular estimation but this is resource intensive and never real-time. A complete and automatic depth prediction pipeline would reduce resources used on this work tremendously.

All these effects can be applied to any film retroactively. Depth maps can serve as a powerful extension for compositing software or as consumer-side post-processing used by a phone or head mounted display (HMD/VR) app. In this way depth maps can be used in automatic 3D augmentation of 2D motion pictures. To demonstrate these capabilities, we have rendered a Virtual Reality 3D stereo experience that uses our framework to generate a 3D illusion. It can be viewed using most headsets including Google Cardboard on YouTube at the following URL: `https://www.youtube.com/watch?v=_v1imUA-dIE`. [6]

2. Related Work

Training on 3D Movie Stereo Pairs has been done in the past as a means of creating new stereo compliments for monocular video. [7] Xie *et al.* previously used stereo pairs to train an end to end network that reconstructed one stereo image from another. Their training method utilized a dataset of approximately 3,000,000 image pairs at a lower resolution than the proposed method. While their model did generate disparity maps at an intermittent step in stereo generation, those depth maps are not ideal for all our explored applications. These maps can be see in Figure 2. Their

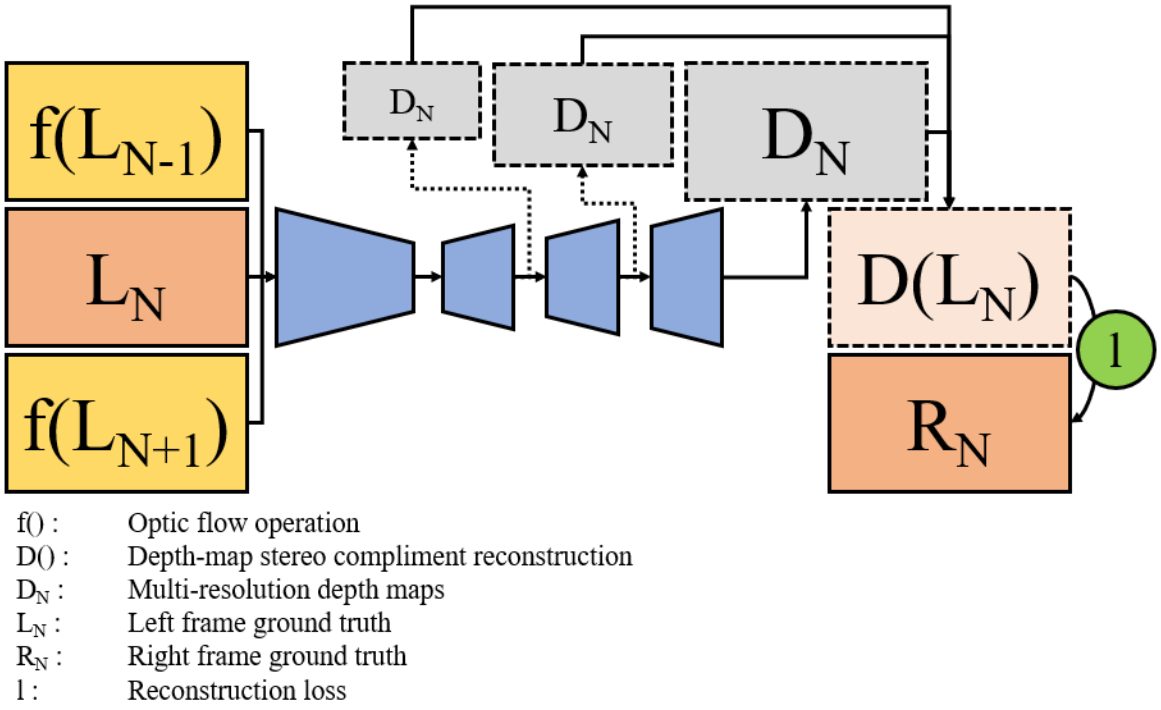

Figure 1. The proposed architecture.

framework would perform insufficiently in the domains of compositing, focus changes, and depth based rendering due to severe banding.

Stereo Reconstruction Loss has been used to successfully produce higher quality depth maps. The idea behind it is that provided a pair of left and right camera views, one view can be constructed from the other using an accurate depth map.

The training method of Godard *et al.* [4] utilized stereo pairs from the KITTI Stereo 2015 self driving dataset. This dataset contains left and right stereo image pairings taken from the top of a car with front facing cameras driving through traffic. [5] The loss function used by Godard *et al.* relied on stereo reconstruction too but instead of using the depth map as an intermittent step like Xie *et al.*, the depth map is the output of the network. This is what makes their network and our network unsupervised. To be unsupervised means that the network's training can not exclusively rely on ground truth data. If complete depth maps were accessible for motion pictures then supervised training could take place. This is not the case, so an unsupervised approach must be used.

This alternative approach, done three years later, provides superior depth maps to Xie *et al.* . The problem with this model is its extreme bias towards automotive footage. This is mostly due to the contents of the KITTI stereo 2015 dataset, but its also exacerbated by the loss function when retraining their network. In Figure 2 there are clear issues when using their trained network on creative scenes. An initial step of our architecture was to re-implement this model and then build upon it.

Title	Frame Count
The Hobbit: An Unexpected Journey	244,105
The Hobbit: The Desolation of Smaug	232,112
The Hobbit: The battle of the Five Armies	207,921
Transformers: The Last Knight	222,448
Transformers: Dark of the Moon	222,104
Prometheus	178,056
Resident Evil: Retribution	137,568
Resident Evil: Afterlife	139,368
Dredd	137,905
47 Ronin	170,786
X-men Apocalypse	207,088

Table 1. Stereo pair frame counts of motion pictures in our dataset.

3. Our Approach

3.1. Architecture

Our approach utilizes aspects of the loss function of Godard et al.[4] and uses a dataset similar to Xie et al.[7] A diagram of the proposed architecture can be seen in Figure 1. The changes we have introduced have permitted for positive and negative disparity, and uses optic flow as input when training and predicting. The network itself uses a Resnet encoder-decoder that has three depth outputs at increasing resolution in succeeding decoding stages. It takes (256 x 640 x 7) inputs. The seven includes two layers of optic flow in the x and y direction between the current frame and

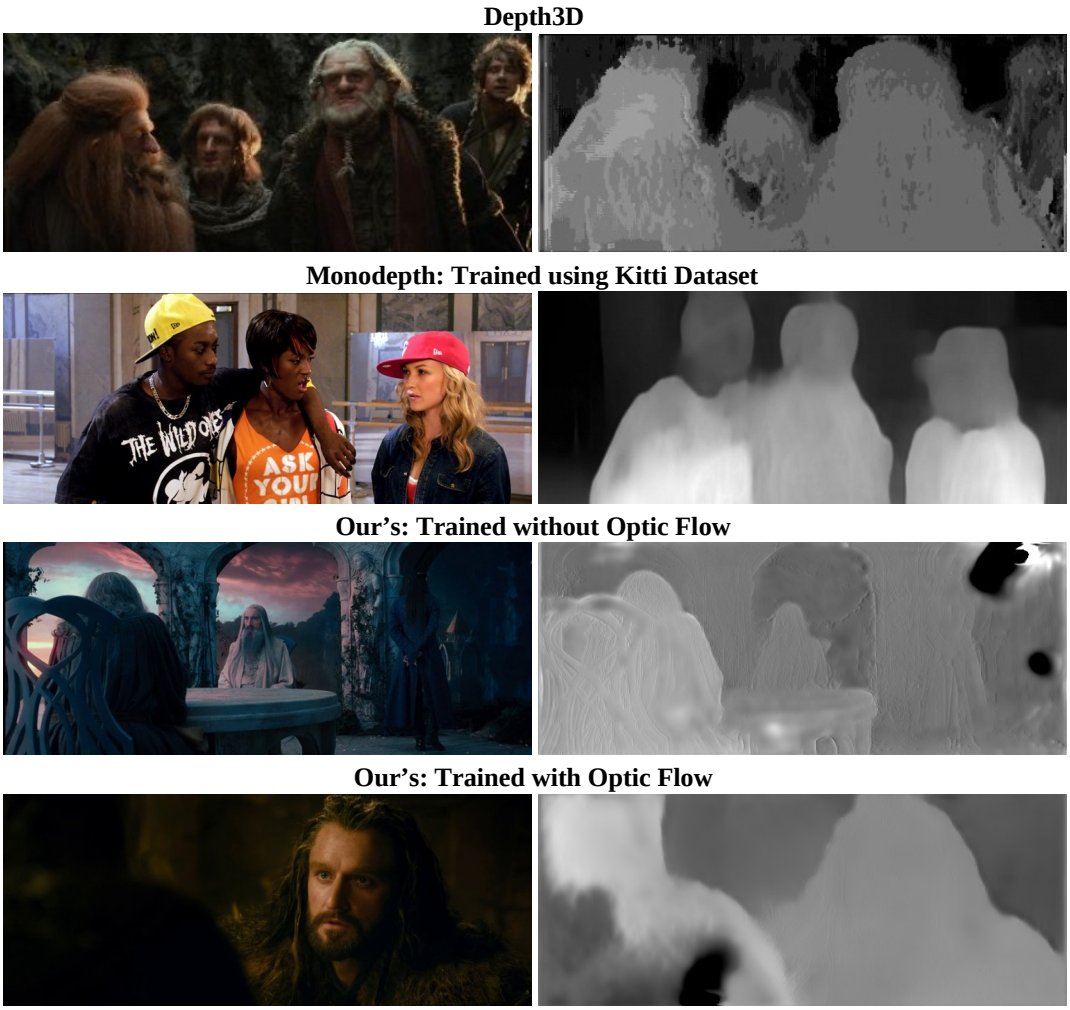

Figure 2. Depth map generation method comparison.

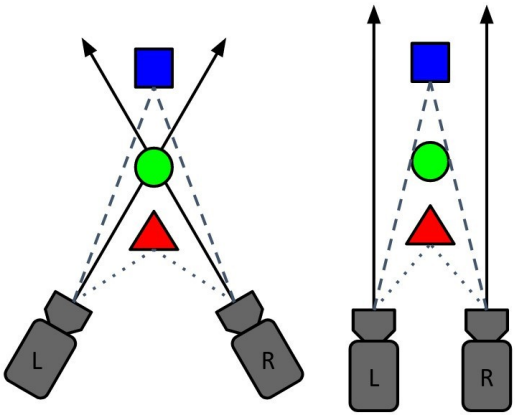

Figure 3. Illustration of disparity in our dataset (left) vs. disparity in Kitti stereo 2015 (right).

previous frame, three color image layers, and two more x and y optic flow between the current frame and next frame.

3.2. Dataset

Our dataset is constructed from 12 3D movies. A breakdown of these films can be seen in table 1. Many movies released today are retroactively converted to 3D through a lengthy unstandardized process called 'dimensionalization' or 'Digital 3D'. In the interest of using real stereo camera information, the films we chose were all shot for 3D (i.e. not dimensionalized or Digital 3D). This choice was made to prevent the network from learning any patterns present in footage artificially made into 3D through 'Digital 3D', and only pick up on real world correspondences between stereo cameras. It should be noted that our 12 films are all live action, and contain a considerable amount if CGI. Films which were completely animated are not included and also already have the means of producing perfect depth maps. The resolution the model was trained at was 640 x 256 pixels. This provides an aspect ratio (screen width / screen height) of 2.5, which is an acceptable middle-ground for most widescreen motion pictures. It

Figure 4. Validation results with optic flow.

also simplifies network structure by being recursively divisible by 2 and therefore does not produce odd numbers when being downsampled by half. Resolution can be increased in future iterations very easily; it is only a matter of more training time.

3.3. Disparity Changes

When initially adapting the architecture of Godard *et al.* , a few issues became apparent immediately. Firstly, the dataset that they originally was trained on only exhibited positive disparity. Disparity is the direction a pixel moves when switching between 3D stereo pairs. Disparity goes in one direction in the Kitti stereo dataset because the two cameras on top of the car were placed facing forward in parallel, converging at infinity. Cameras on a 3D motion picture set do not do this and instead converge on a subject. This subject is most frequently an actor in the scene. What this means is that a creative scene's disparity map must have positive and negative disparity to completely describe the difference between the left and right images. An illustration of why this happens can be seen in Figure 3. Our first change made to the architecture was an adjustment to its loss function that allowed for disparity in both directions. Without making this change issues like those observed in Appendix A occur. There you can see predictions for elements in front of the point of convergence are accurate but those behind are completely wrong.

3.4. Frame Triplets

When inputting frames into the network individually while accounting for disparity in both directions, results showed marked improvement over the network trained on automotive footage. In an effort to increase performance further, triplets of frames were put into the network. In other words, instead of inputting a single image, we would pass in the image before it and the image after it in the motion picture.

Initial results from this implementation were sub-satisfactory

and were worse than when input was individual frames. The reason behind this is partially obfuscated due to the network's black box properties, but the basic principle is likely that the input arrays for each pixel are not 'aligned' when representing three different images. This issue can be rectified using alternative representations like optic flow.

3.5. Optic Flow

OpenCV has a function cv.calcOpticalFlowFarneback() that uses the method outlined by Farneback in 2003.[2] It provides a dense optic flow approximation for every pixel in a given frame. We calculated this at a 720p resolution before down-sampling it to 640 x 256 in order to maximize its accuracy. The augmentation of input data with optic flow has shown to improve depth estimation. This can be seen in Figure 2. Optic flow matrices were scaled to have an absolute average of 1.0 flow, as we are only interested in the intra-frame differences in flow when predicting depth. A value of 1.0 in an optic flow frame would represent average positive flow in that current frame, 0 would represent no change, and -1.0 represents flow of equal magnitude in the opposite direction. The flow of the frame before and after are included for every input. Optic flow in the x and y direction are stored in separate matrices.

4. Results

Before implementation of the novel optic flow input, results already were showing successfully adaptation to the new dataset. The loss function required minor reworking before the network was able to produce good maps. An issue however was a tremendous amount of 'noise' and 'spike' artifacts. These are most clearly seen in Figure 2. Spikes were mitigated by introducing limits to the disparity which prohibited exaggerated spikes of impossible disparity. Depth noise was reduced by adding the optic flow frames. After the introduction of optic flow results were more 'smooth'. A smother depth map is more true to reality and makes reconstruction less jarring.

Our ultimate implementation that took optic flow input has only been trained on the 'The Hobbit: Desolation of Smaug'. Calculating optic flow for every frame needs to be done once before training in this way. For us it took about 5 hours to complete a single film using a Virtual Machine allocated 8 cores of an Intel E5-2697 Xeon CPU and 32GB of RAM. Despite this limited training, inferences can be gleamed from the results. Training itself can take 72 hours using an Nvidia Quadro P5000.

In order to assess our network with limited training a specific test film was chosen. 'Lord of the Rings: the Fellowship of the Ring' is a 2D motion picture with a lot of overlapping subject matter with the Hobbit. Looking at the depth results from this we can see a few patterns. Looking at Figure 4, there is notably better performance exhibited on subjects that are within both films, this includes elves and dwarfs. The shots in the test film are nevertheless unique in terms of composition, so it appears that the network is able to learn to recognize subjects on some level to help it infer depth. With additional training on more subject matter it's possible that issues with other subjects may be resolved. For instance, if the network is exposed to transformers, it is possible a mechanic looking creature like Sauron will be easier to predict as it will be forced to generalize.

5. Conclusion

Based on the results seen here there is clear promise in the potential application of this technology for motion pictures. With very little resources and at almost no cost this framework was assembled and made to work by an individual undergraduate student. Improving upon this dramatically can be done with the introduction of multiple stages of training or the use of supplemental synthetic data. These results show no signs of hitting a barrier and are limited only by the work invested in their refinement.

References

[1] B. Atapour-Abarghouei. Real-time monocular depth estimation using synthetic data with domain adaptation via image style transfer. *CVPR*.

[2] Farneback. Two-frame motion estimation based onpolynomial expansion. *Computer Vision Laboratory*,, 2003.

[3] M. Firman. RGBD Datasets: Past, Present and Future. In *CVPR Workshop on Large Scale 3D Data: Acquisition, Modelling and Analysis*, 2016.

[4] F. B. Godard, Aodha. Digging into self-supervised monocular depth estimation. 2019.

[5] M. Menze and A. Geiger. Object scene flow for autonomous vehicles. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[6] O. Thompson. Ai predicted stereo vr auto-remaster: Test 3.

[7] F. Xie, Girschick. Deep3d: Fully automatic 2d-to-3d video conversion with deep convolutional neural networks.2016.

Appendix A. Disparity Issues

Figure 5. Results when training with disparity only being in one direction.