

**A Novel Convolutional Neural Network GSDP-GridNet:
for Geometric & Semantic Consistent Depth Prediction
for 3D Ken Burns Synthesis from a Single Image**

A Dual Degree Project Report

submitted by

PRATEEK SONI

under the guidance of

PROF. MANSI SHARMA

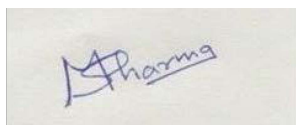


**DEPARTMENT OF ELECTRICAL ENGINEERING
INDIAN INSTITUTE OF TECHNOLOGY MADRAS.**

June 2021

THESIS CERTIFICATE

This is to certify that the thesis titled **A Novel Convolutional Neural Network GSDP-GridNet: for Geometric Semantic Consistent Depth Prediction for 3D Ken Burns Synthesis from a Single Image** , submitted by **Prateek Soni (EE16B145)**, to the Indian Institute of Technology, Madras, for the award of the degree of **Bachelors of Technology** and degree of **Masters of Technology**, is a bonafide record of the research work done by him under my supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.



Prof. Mansi Sharma

Research Guide

Inspire Faculty

Dept. of Electrical & Engineering

IIT-Madras, 600036

Place: Chennai

Date: June 2021

ACKNOWLEDGEMENTS

Firstly, I would like to express my sincere gratitude towards my advisor, Dr. Mansi Sharma. She gave me the freedom to work and explore any topic but always helped me and guided me towards the right direction. Without her invaluable guidance, insights, time, and support, this work would not have been possible.

I would also like to thank all the professors in the Electrical Engineering Department as well as other departments whose courses I have taken and learned a lot from, which has been put to use in this thesis.

ABSTRACT

KEYWORDS: Ken Burn Effect, Depth Maps, View Synthesis, Inpainting, GridNet.

3D Ken Burns effect is an image processing that mimics real camera motion to create a video from a static picture. This photographic effect consists of the combination of tracking shots, panning and zooming effect. Generating 3D ken burns effect from a single image is a challenging task that takes a lot of time and requires advanced editing skills. This effect is created by generating 3D scene of the initial image and then rendering novel views along the camera path.

We propose a fully automated solution to generate a 3D ken burn effect from a single image in real-time. We divide this method in mainly two components: depth prediction and image inpainting for view synthesis. We propose a novel GridNet based convolution neural network for predicting geometric and semantic consistent depth maps that are suitable for view synthesis. We also propose a UNet based depth refinement network for up-scaling and refinement of the predicted depth and to achieve accurate depth prediction at object boundaries. We then introduce our view synthesis pipeline to generate the 3D ken burn effect from the predicted depth and the input image. We convert the image and predicted depth into a point cloud and render novel views from different camera positions. We perform joint color and depth inpainting to fill the disocclusion in the rendered novel view. We use inpainted depth to extend the point cloud and repeat the process until the point cloud is sufficiently extended. The proposed depth prediction model achieves state-of-the-art performance in both qualitative and quantitative evaluation on iBims and NYUv2-Depth dataset. Our subjective analysis shows that the proposed fully automated 3D ken burn effect method achieves better results compared to the manual and existing methods of generating 3D ken burns effect.

TABLE OF CONTENTS

THESIS CERTIFICATE	i
ACKNOWLEDGEMENTS	ii
ABSTRACT	iii
LIST OF TABLES	vi
LIST OF FIGURES	vii
1 Introduction	1
1.1 The Ken Burn Effect	1
1.2 Depth Estimation	2
1.3 View Synthesis and Image Inpainting	2
1.4 Contribution	3
2 Related Work	4
3 3D Ken Burn Effect View Synthesis	6
3.1 Geometric & Semantic Consistent Depth Prediction	6
3.1.1 Depth Estimation	6
3.1.2 Depth Up-scaling and Refinement	8
3.1.3 Depth Smoothing	9
3.2 View Synthesis and 3D Inpainting	9
4 Training and Experiments	12
4.1 Training Data & Experimental Settings	12
4.2 Depth Comparison	13
4.3 View Synthesis Comparison	14
4.4 Subjective Quality Analysis	14
5 Results	17

5.1	Depth Results	17
5.2	View Synthesis Results	20
6	Conclusion	23

LIST OF TABLES

4.1	Comparison of different methods on NYUv2-Depth data.	13
4.2	Comparison of different methods on iBims-1 data.	13
4.3	View Synthesis quality analysis	14

LIST OF FIGURES

3.1	Depth Estimation Network	7
3.2	Depth Refinement Network	8
3.3	View Synthesis Model	10
4.1	Results from subjective user study comparing Niklaus et al, Ours and 2D ken burn	16
5.1	Comparison of Depth maps on Deer Scene.	17
5.2	Comparison of Depth maps on Hall Scene.	18
5.3	Comparison of Depth maps on Baba Scene.	18
5.4	Comparison of Depth maps on NAC Canopy Scene.	19
5.5	Comparison of Depth maps on NAC Trees Scene.	19
5.6	Comparison of Depth maps on Jamuna bus stand Scene.	20
5.7	Snapshots of 3D ken burns effect of Baba Scene.	20
5.8	Snapshots of 3D ken burns effect of Deer Scene.	21
5.9	Snapshots of 3D ken burns effect of Jamuna bus stand Scene.	21

CHAPTER 1

Introduction

1.1 The Ken Burn Effect

The Ken Burns effect is a photographic effect that consists of the combination of tracking shots, panning and zooming effect, created from still imagery. A tracking shot is a camera movement during the recording. It can be lateral to the scene, following the motion of a character. Or it could be orthogonal, in that case the camera moves either backward or forward to the scene and creates a realistic zoom as if the spectator was getting farther or closer. Existing image- and video-editing tools could be used to freely augment photographs with virtual animation and depth effect, *i.e.*, enabling motion parallax as the virtual camera scans over a still picture. This visual effect is called as 3D Ken Burns effect. Compared to its traditional 2D counterpart, creating the illusion of 3D movement adding 3D zoom, pan and parallax transitions allows much more compelling experiences. It is widely used in order to embed still images into videos. This 3D effect has become increasingly popular in films, documentaries, and commercial media production where they lack video footage to illustrate the words. However, creating such effects from a single photograph fully automatically is a challenging task that takes a lot of time and requires advanced editing skills.

For Generating 3D Ken Burns effect 2 cropping windows are chosen in the image. The first window would be the starting frame and the second window would be the ending frame of the video. The automated ken burns effect synthesis chooses the start and ending frame automatically such that it minimizes the disocclusions. The in-between frames of the video are created by a view synthesis model which converts the image and its depth map into a 3d point cloud and then render the novel views by moving the virtual camera along the two endpoints. View synthesis model also performs joint color and depth inpainting in the dis-occluded regions and extends the 3d point cloud in the process, this ensures that rendered views are free from disocclusion artifacts.

1.2 Depth Estimation

As mentioned in the above section, we need depth maps to create 3d point cloud, required for rendering novel views. The problem statement is to synthesis ken burn effect from single image and so depth estimation should also be achieved from single image. The task of estimating depth map from a single image is termed as monocular depth estimation. For each pixel of the image a depth value is assigned which is an estimate of the distance between the camera and object in the scene. Especially for synthesizing ken burn effect we need geometric and semantic consistent depth maps with no depth boundary error ideally. To achieve this we introduced a novel GridNet [Fourure *et al.* (2017)] based convolutional neural network that make explicit use of geometric information from bilateral geometric map of [Mansi *et al.* (2020)] and semantic information from pool_4 layer of pre-trained VGG-19 network [Simonyan and Zisserman (2014)]. To achieve consistent depth of each object in the depth map we make use segmentation maps from [Vladimir Nekrasov (2018)] and attribute average depth of the mask to every pixel in that mask in the depth map. 3D ken burns effect synthesized using our depth maps shows more pleasing to the eyes results.

1.3 View Synthesis and Image Inpainting

Image Inpainting is the growing field in the past few years, especially with rise of deep learning. It can be done using convolutional neural network based deep learning methods like [Liu *et al.* (2018a)] or more recently with generative models like [Yeh *et al.* (2017)] and [Xie *et al.* (2012)]. Image inpainting is mostly used for photo editing purpose like restoring old pictures, CT-scan or MRI images.

In view synthesis when we move the virtual camera along the specified path to generate novel views it causes some disocclusion. These disocclusions needs to fixed with geometric consistent image inpainting for generating 3d ken burn effect with distortions. We use a convolutional neural network inspired from GridNet [Fourure *et al.* (2017)] to make image inpainting model. This model is used to fill the missing information caused of dis-occluded areas in the rendered novel views.

1.4 Contribution

In this work, we proposed a fully automated solution to generate a 3D ken burn effect from a single image in real-time. To achieve this we divide the work into two main components: depth prediction and image inpainting for view synthesis. We proposed a novel GridNet based convolution neural network for predicting geometric and semantic consistent depth maps. We also proposed a UNet based depth refinement network for up-scaling and refinement of the predicted depth and to achieve accurate depth prediction at object boundaries. We worked on improving existing view synthesis methods to generate the 3D ken burn effect from the predicted depth and the input image. And we finally showed from our informal user study that we achieve better results compared to the manual and existing methods of generating 3D ken burns effect.

CHAPTER 2

Related Work

The novel view synthesis of scenes or 3D objects from reference images captured from different viewpoints has a wide range of applications in image or video manipulation [Klose *et al.* (2015)], virtual and augmented reality [Hedman *et al.* (2017)], and 3D displays [Didyk *et al.* (2013)]. Existing depth-image-based rendering (DIBR) or computer vision based techniques generate high-quality view synthesis results [Hedman and Kopf (2018)]. With the advent of deep neural networks, learning-based approaches have become steadily gaining popularity for novel view synthesis [Flynn *et al.* (2016)]. Typically, most of the view synthesis algorithms require multiple input views acquired from sparse camera viewpoints to render a virtual view. Though, our target in this paper is to generate novel views with a virtual camera scan along a path and zoom, given only a single monocular image.

There are inherent challenges of rendering novel viewpoints from a single image. Existing learning based methods are giving quality output only for specific scene types, 3D object models, domain specific data such as light fields [Srinivasan *et al.* (2017)]. The quality of view synthesis is dependent on precision of scene geometry. Several approaches for depth estimation from a single image have been presented based on normal maps [Liu *et al.* (2018b)] or layered depth [Tulsiani *et al.* (2018)]. In proposed formulation, we predict scene depth suitable for rendering high quality views for synthesizing a plausible 3D Ken Burns effect. We particularly present our depth prediction and view synthesis results on challenging cases such as natural scenes with moderate or large depth variation, natural lighting, reflections and transparency, high details or thin structures. In such complex scenarios, the existing depth estimation approaches usually suffer to customize the depth prediction in suppressing geometric artifacts for the task of view synthesis.

Synthesizing realistic visual camera effects such as bullet-time freeze effect, zoom in/out, 2D-to-3D, depth-of-field synthesis, photo pop-up for creative pursuits can be produced using 3D scene geometry or scene layout information. We focus on the more

challenging 3D Ken Burns effect which is a camera motion effect with realistic zoom, pan and parallax. Previous efforts tried to create fly-through effects from a single image. [Horry *et al.* (1997)] semi-automatic system demand user intervention in foreground object segmentation and generating a mesh based simplified representation of the scene. The 3D illusion effect of camera flying is created by projecting images on the mesh geometry. Follow-up work improves the scene representation to accommodate diverse camera motions and pictures containing multiple vanishing points. Still realistic effects cannot be achieved for general scenes and the process is not fully automatic, demands significant manual supervision. [Zheng *et al.* (2009)] method somewhat sort out these issues and output realistic parallax from still pictures, but it works only with multiple images as input. 3D ken burns [Niklaus *et al.* (2019)] approach is close to our work, animating still views with a virtual camera scan adding parallax and compelling results on different scene types including “indoor”, “landscape”, “outdoor”, and “portrait”. However, in many challenging natural scenes with high details or thin structures, occlusions, reflection or transparency, their approach gives unnatural synthesis results with unexpected blurring, flickering around object boundaries, and semantic distortions. Our approach synthesizes desirable effects in such complex scenes with realistic motion parallax and strong depth perception.

CHAPTER 3

3D Ken Burn Effect View Synthesis

For ken burn effect we need 2 cropping windows, one as start frame and other as end frame for creating the virtual zoom. Instead of asking user to provide coordinates of these 2 windows our algorithm chooses the start and ending frame automatically such that it minimizes the disocclusions. The framework for generating 3D ken burn effect consist of two main components, depth estimation and view synthesis.

3.1 Geometric & Semantic Consistent Depth Prediction

The Ken burn effect is to be performed on the single image and the depth estimation should also be achieved from the single image. Unsupervised learning methods can perform the depth estimation but their performance is still behind the supervised learning methods. Therefore we choose to do depth estimation task using supervised learning. Inspired from Niklaus et al. [Niklaus *et al.* (2019)] work we used two separate neural networks, one for coarse depth estimation and one for up-scaling and depth refinement.

3.1.1 Depth Estimation

We introduced a GridNet architecture [Fourure *et al.* (2017)] that explicitly uses the semantic and geometric information of the scene to predict the depth from a single image. To extract the semantic features of the image we pass the input image to a pre-trained VGG-19 [Simonyan and Zisserman (2014)] network and extract the features from the pool_4 layer of it. To extract the geometric features of the scene we make use of the bilateral geometric map from [Mansi *et al.* (2020)] and pass it through a pre-trained VGG-19 [Simonyan and Zisserman (2014)] network and extract the features from the pool_4 layer of it. Providing this information explicitly to our model helps us in achieving geometric and semantic consistent depth prediction.

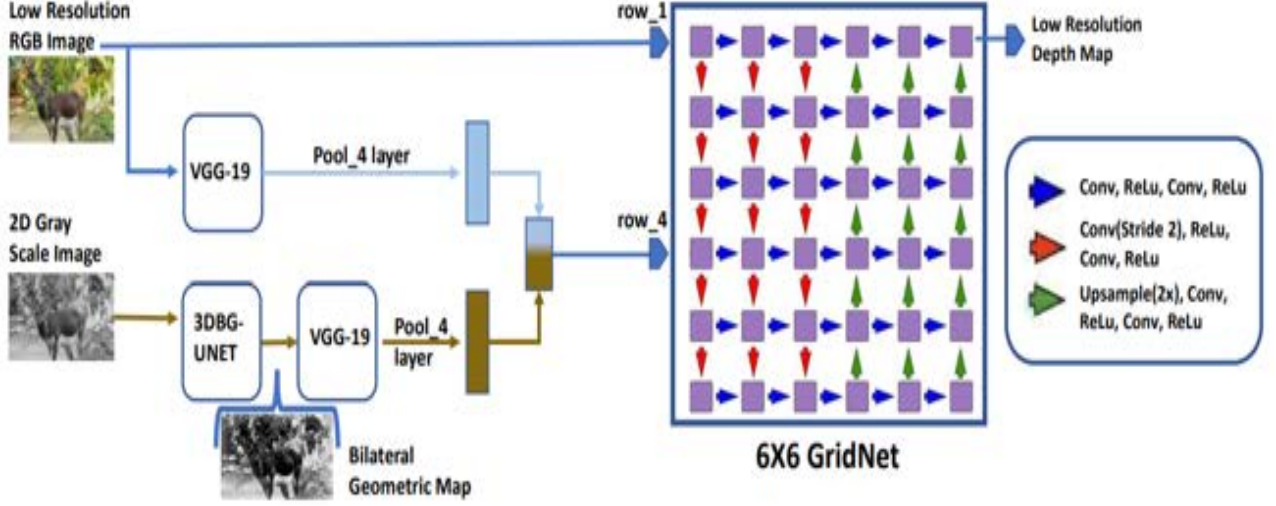


Figure 3.1: Depth Estimation Network

We choose a GridNet architecture of 6 rows and 6 columns where the first 3 columns perform downsampling and the last 3 columns perform upsampling with a per-row channel size of 32, 64, 128, 256, 512, and 512. We give the input image into the first row and explicitly provide the semantic and geometric features to the 4th row of grid. This information encourages the network to better capture the geometry of the scene and thus solve the problem of geometric distortion. Since the rows followed after we provide this information have a relatively higher number of channels it further encourages the network to make use of the semantic and geometric information.

Loss function: To supervise the depth estimation network we make use of scale invariant \mathcal{L}_1 data loss and multi-scale level MSE gradient loss. The data loss penalizes the network to produce depth maps that are different from the ground truth, pixel-wise.

$$\mathcal{L}_{data} = \frac{1}{WH} \sum_{u,v} (\xi_N(u,v))^2 - \left(\frac{1}{WH} \sum_{u,v} \xi_N(u,v) \right)^2$$

where $d_N(u,v)$ is the ground truth depth, $\hat{d}_N(u,v)$ is the estimated depth at pixel (u,v) , and $\xi_N(u,v) = \log(d_N(u,v)) - \log(\hat{d}_N(u,v))$

The gradient loss penalizes the gradient of the depth estimation and gradient of the ground truth to be different. This loss function enforces smoothness in the homogeneous regions of predicted depth map.

$$\mathcal{L}_{grad} = \frac{1}{WH} \sum_{k=1,2,4,8,16} \sum_{u,v} \|\nabla^k \hat{d}_N(u,v) - \nabla^k d_N(u,v)\|^2$$

The gradient of the image is computed by shifting the image by one pixel horizontally (or vertically) and taking the difference between the intensities. $\nabla_x I(u,v) = I(u,v) - I(u-1,v)$ at scale 1, similarly for scale $k \geq 1$: $\nabla_x I(u,v) = I(u,v) - I(u-k,v)$.

We give more emphasises on gradient loss than data loss while training the depth estimation network and present the total loss function for network as

$$\mathcal{L}_{depth} = \mathcal{L}_{grad} + 0.0001\mathcal{L}_{data}$$

For training we chose the Adam optimizer with $\alpha = 0.0001$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$ and trained the network on 20k samples from NYUv2 depth dataset [Nathan Silberman and Fergus (2012)], 8k samples MegaDepth dataset [Li and Snavely (2018)] and 4k samples DIML [(DIML)] dataset for 200K iterations.

This depth estimation is not done directly on the full scale image. Instead, the input image is resized so that its largest dimension is 512 pixels. Therefore an other step is necessary in order to produce a full scale depth map.

3.1.2 Depth Up-scaling and Refinement

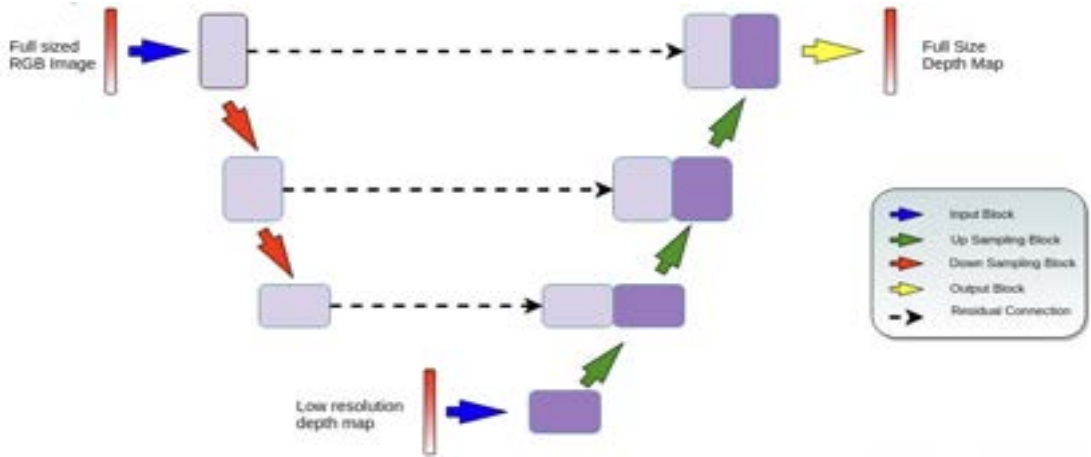


Figure 3.2: Depth Refinement Network

We use the another neural network to perform depth refinement. The refinement

network is a modified UNet architecture that takes the original image as the input, down-scales it to the same size as the size of above depth estimate so it can be inserted in the network and then up-scaled and refined with the guidance of original input image. This model is also trained using the same loss functions as for the depth estimation. This gives us the full size geometric & semantic consistent depth map required to synthesize 3D ken burn effect.

For training we chose the Adam optimizer with $\alpha = 0.0001$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. For training data, we downsample and distort the ground truth depth to simulate coarse depth and use the ground truth depth and downsampled depth for training. We took 8k samples from NYUv2 depth dataset [Nathan Silberman and Fergus (2012)], 5k samples MegaDepth dataset [Li and Snavely (2018)] and 3k samples DIML [(DIML)] dataset for 100K iterations.

3.1.3 Depth Smoothing

Often the depths are assigned inconsistently inside the region of the same object and it results in strange view synthesis outputs such as objects being stretched (or torn apart) or objects sticking to another objects. To overcome these issue we locate the objects in the image and assume that the depth is constant on these objects. We make use of pre-trained Mask R-CNN by Light Weight RefineNet [Vladimir Nekrasov (2018)] to generate segmentation masks of different objects in the scene. The average depth on a mask is then attributed to every pixel in that mask in the depth map.

Our depth maps are semantic-aware and preserve high frequency details in the depth map needed for artifact free 3D ken burn view synthesis. Ken Burn effect generated using depth map from our depth prediction pipeline shows geometrically and semantically consistent results and are pleasing to the eyes.

3.2 View Synthesis and 3D Inpainting

The View Synthesis is the next important step to generate 3D-ken burn effect after estimating Geometrically and Semantically consistent depth of the image. View synthesis is the process to generate new scenes of an image from a shifted camera position. Some

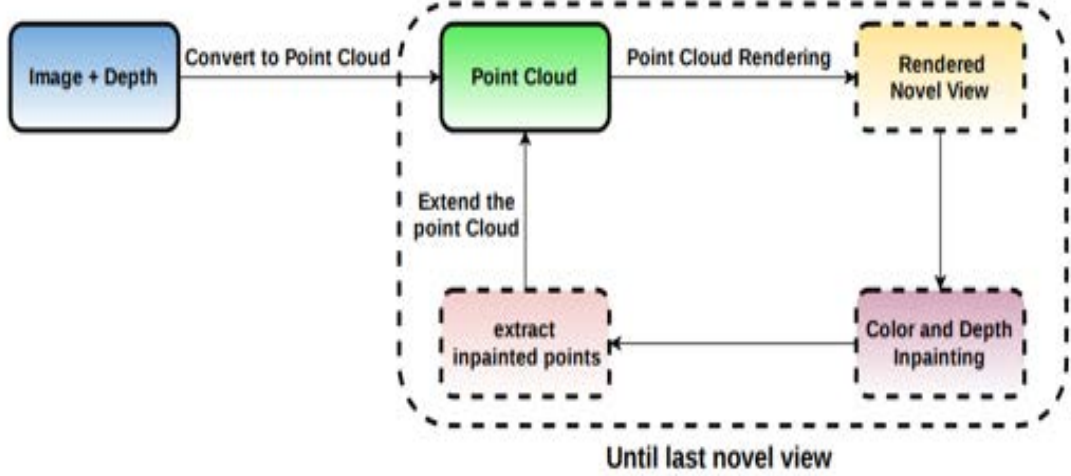


Figure 3.3: View Synthesis Model

parts of the object may not be visible from the starting camera position that can become visible when the scenes are generated from different camera position, these regions were occluded in the initial scene. To address the problem of disocclusion while generating new views it is necessary to fill the missing information in these regions to create the realistic 3D effect, the process of filling this information is known as image inpainting.

Off the shelf color inpainting methods fails to perform geometrically consistent inpainting. To perform inpainting in which the filled-in area resembles the background with the clear separation of the foreground object we need inpainting methods which incorporate depth. There are many techniques for color inpainting and some of them do take depth information into account. But only few techniques focuses on the problem of joint color and depth inpainting. We performed joint color and depth inpainting using a single neural network specifically trained to fill disocclusion.

Architecture: We used a architecture similar to our depth prediction network, we used a GridNet architecture that takes color and depth of the incomplete rendered novel view and returns the inpainted color and depth. We choose a GridNet with 4 rows and 4 columns with channels size of 32, 64, 128, and 256 in the 4 rows.

Loss function: To supervise the color inpainting we use the reconstruction and perceptual loss. Reconstruction loss enforces the match between the reconstructed image and the ground truth image. It is been observed that \mathcal{L}_2 loss penalize less to the small variations compared to \mathcal{L}_1 loss and produces blurry outputs. Hence, we used pixel-wise

\mathcal{L}_1 loss for reconstruction.

$$\mathcal{L}_{rec} = \|I - \hat{I}\|_1$$

where \hat{I} is the reconstructed novel view and I is the ground truth novel view.

For perceptual loss we used a pre-trained VGG-19 [Simonyan and Zisserman (2014)] to extract features from the ground truth and inpainted novel view. Perceptual loss penalize the inpainting network for producing reconstruction with different features maps than the ground truth one.

$$\mathcal{L}_{perc} = \sum_{k \in \mathcal{K}} \frac{1}{N_{\phi_k}} \|\phi_k(I) - \phi_k(\hat{I})\|_1$$

where ϕ_k is the feature map after the k th convolution layer in the VGG-19 network, N_{ϕ_k} is the size of the feature map ϕ_k and \mathcal{K} is the set of first 4 pooling layers of VGG-19 network.

To supervise the depth inpainting we use the scale invariant \mathcal{L}_{data} data loss and \mathcal{L}_{grad} gradient loss function same as the loss function in depth estimation network. Now, the total loss function to supervise the training of joint color and depth inpainting becomes:

$$\mathcal{L}_{inpaint} = \mathcal{L}_{rec} + \mathcal{L}_{perc} + \mathcal{L}_{grad} + 0.0001\mathcal{L}_{data}$$

For training we chose the Adam optimizer with $\alpha = 0.0001$, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We trained our network multi novel view synthetic dataset of [Niklaus *et al.* (2019)] for 100k iteration with learning rate of $10^{(-5)}$ and learning rate decay of 0.9999.

Now, In order to do view synthesis we first create a point cloud from the input image and the estimated depth. Then we use point cloud to render consecutive novel views from new positions along the specified camera path. However, the render views faces the issue of disocclusion since the point cloud is only a partial view of the scene. We perform geometrically consistent joint color and depth inpainting to fill-in the missing information in the occluded regions to get the complete novel view from the incomplete render. We extract the inpainted points from the reconstructed novel view and extend our point cloud in each iteration until we get to the last novel view.

CHAPTER 4

Training and Experiments

In this section, we describe experiments to evaluate the performance of proposed GSDP-GridNet neural network with the state-of-the-art CNN algorithms for monocular depth estimation and further how that depth effects affect the view synthesis results.

4.1 Training Data & Experimental Settings

We used benchmark NYUv2 RGB-D dataset [Nathan Silberman and Fergus (2012)], MegaDepth dataset [Li and Snavely (2018)], and DIML RGB-D dataset [(DIML)] for training and validation of our novel depth prediction method. NYUv2 dataset consists of indoor scenes of high quality aligned RGB and depth images acquired from a Microsoft Kinect sensor. Where as Megadept and DIML dataset brings more scenes into training as MegaDepth dataset consist of outdoor scenes and DIML dataset consists of both indoor and outdoor scenes. The comparative analysis is performed on 2 dataset separately and presented in the next section. We used iBims-1 benchmark [Koch *et al.* (2018)] dataset which contains 100 images and 450 images from the NYUv2-Depth dataset [Nathan Silberman and Fergus (2012)].

We implemented our depth and view synthesis pipeline in PyTorch. The models are trained on Intel i7-9750H CPU, 16 GB RAM, RTX 2080 GPU with 8 GB VRAM with GPU acceleration. The details of the loss functions, optimizer and hyper-parameters settings is written in the above chapter [3]. Training for depth estimation network took 2+ days. The depth refinement model training took around 12+ hours. The training of view synthesis network took around 20+ hours.

Our depth estimation pipeline takes 2-3 seconds to estimate the depth map of a image. The view synthesis pipeline takes around 7-8 seconds per image to synthesize a 3D ken burn video from depth map and input image. Both depth and view synthesis pipeline combine takes 9-10 seconds to generate 3D ken burn effect video.

4.2 Depth Comparison

We compare our depth estimation GSDP-GridNet with state-of-the-art monocular depth estimation methods, including Niklaus et al. [Niklaus *et al.* (2019)], DenseDepth [Al-hashim and Wonka (2018)], FCRN [Laina *et al.* (2016)], 3DBGES-UNet [Mansi *et al.* (2020)] and SharpNet [Ramamonjisoa and Lepetit (2019)].

We perform quantitative analysis on iBims-1 benchmark [Koch *et al.* (2018)] dataset and NYUv2 Depth dataset [Nathan Silberman and Fergus (2012)]. We evaluate our model using most common error metrics, root mean squared error (RMSE), average error (\log_{10}), average relative error (rel), threshold accuracies (σ) on both iBims-1 benchmark and NYUv2 dataset. Additionally, we use mean structural similarity index (mSSIM) and depth reliability metric (DERM) while evaluating on NYUv2 dataset.

Table 4.1: Comparison of different methods on NYUv2-Depth data.

	SharpNet	DenseDepth	FCRN	3D Ken Burns	GSDP-GridNet
rel ↓	0.14	0.12	0.13	0.10	0.12
\log_{10} ↓	0.05	0.05	0.05	0.04	0.09
RMS ↓	0.49	0.46	0.57	0.36	0.32
σ_1 ↑	0.89	0.85	0.81	0.90	0.93
σ_2 ↑	0.98	0.97	0.95	0.98	0.99
σ_3 ↑	0.99	0.99	0.99	0.99	0.99
mSSIM ↑	0.72	0.98	0.95	0.95	0.99
DERM ↑	0.51	0.72	0.64	0.70	0.80

Table 4.2: Comparison of different methods on iBims-1 data.

	3DBGES-UNet	DenseDepth	FCRN	3D Ken Burns	GSDP-GridNet
rel ↓	0.23	0.13	0.13	0.10	0.11
\log_{10} ↓	0.04	0.04	0.04	0.04	0.03
RMSE ↓	0.34	0.46	0.47	0.47	0.22
σ_1 ↑	0.46	0.59	0.53	0.90	0.92
σ_2 ↑	0.73	0.95	0.83	0.97	0.97
σ_3 ↑	0.90	0.99	0.97	0.99	0.99

4.3 View Synthesis Comparison

We generate high-quality novel view synthesis results using geometric and semantic consistent depth maps obtained from the GSDP-GridNet model. Intermediate frames from rendered 3D ken burns effect animation when the zoom is maximum are shown in Fig. 5.7, Fig. 5.8 and Fig. 5.9 for the purpose of comparison. These videos are synthesized by shifting virtual camera positions using the estimated depth maps of 3DBGES-UNet [Mansi *et al.* (2020)], FCRN [Laina *et al.* (2016)], DenseDepth [Alhashim and Wonka (2018)], 3D Ken Burns [Niklaus *et al.* (2019)], and our proposed GSDP-GridNet. To clearly see the generated 3d ken burn effect using different methods, readers are encouraged to see the results uploaded at the website [Prateek and Mansi (2021)] or by clicking [here](#).

For quantitative comparison of the synthesized 3D ken burns effect we compute blinds scores for each of the them. Blinds scores considers distortion, quality and artifacts in the video. There are 4 metrics are Brisque, Nique, Pique and Blinds.

Table 4.3: View Synthesis quality analysis

	3DBGES-UNet	DenseDepth	FCRN	Kens Burn	GSDP-GridNet
Brisque ↓	40.26	39.74	39.73	39.43	39.33
Nique ↓	3.66	3.65	3.67	3.66	3.64
Pique ↓	44.48	44.46	44.87	44.69	44.38
Blinds ↑	6.00	5.98	5.97	6.02	6.03

4.4 Subjective Quality Analysis

We carry out an informal user study to evaluate the performance of our proposed method for creating the 3D Ken Burns effect. The five complex scenes are selected from the 3D HDR stereo dataset Wadaskar *et al.* (2019). The scenes are chosen from indoor, outdoor, nature and portrait. This covers a variety of complex scenarios. The subjective analysis is performed with the help of eight participants from IITM 'Computational Imaging & Display' lab. We compare the proposed approach for creating the 3D Ken Burns effect with Niklaus et al. [Niklaus *et al.* (2019)] and 2D ken burns effect.

To evaluate the impact of geometric, semantic, and inaccurate depth boundaries distortions on rendered videos, we carried out in-lab subjective experiments. The conducted visual perception experiments are basically designed to observe the realism of Ken Burns effect on different chosen scenes. The age distribution of eight subjects (3 female, 5 male) participated in the study ranged between 20 to 35. The major factors that influence visual quality in rendered Ken burns videos that affect human realism ratings are identified as a) Geometric distortions, b) Semantic distortions, c) unreliable depth boundaries. The source of such artifacts could be visibility, disocclusion, resampling issues associated with 3D warping and distortions caused by rendering virtual motion parallax effect. The subjects were asked two main questions: 1) What are the key factors you perceive that influence the realism of synthesized Ken Burns effect?. 2) How do you perceive the variations in different distortion factors that affect your judgment of realism?. The participants were asked to rate each view synthesis results based on visual quality and 3D perception on a continuous quality scale with five levels ranging from 1 – 5: 1 (poor), 2 (fair), 3 (good), 4 (very good), 5 (excellent). The videos are presented to the subjects randomly. The time for each evaluation is limited to 20-25 sec and every rendered scene is evaluated 5+ times. The rating or “opinion score” provided for each synthesized sequence is recorded and the mean opinion score (MOS) is computed. The MOS (mean opinion score) for all the five scenes are plotted in Fig 4.1.

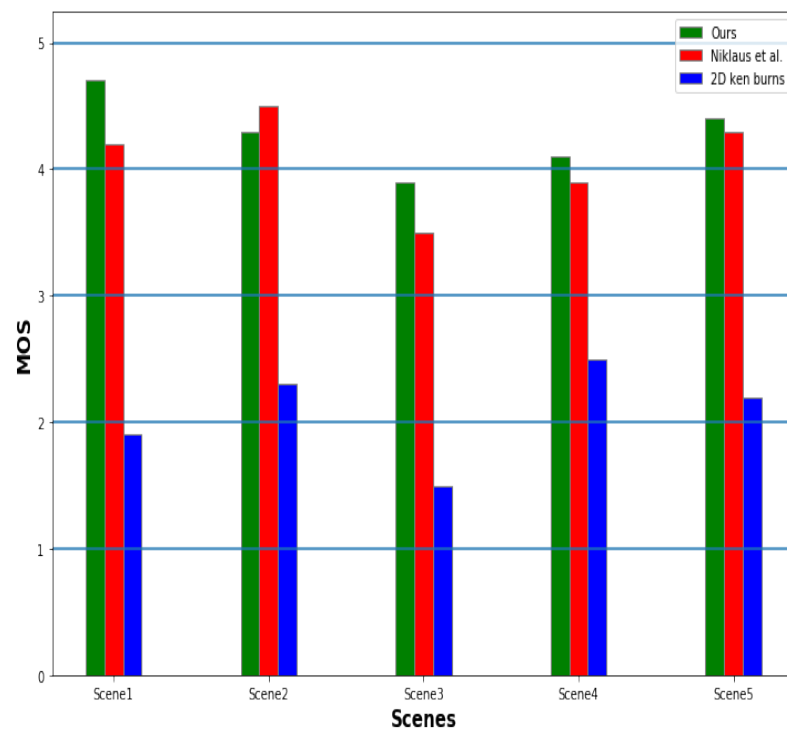


Figure 4.1: Results from subjective user study comparing Niklaus et al, Ours and 2D ken burn

CHAPTER 5

Results

5.1 Depth Results

The depth estimation results are shown in the below figures on complex scenes selected from 3D HDR database [Wadaskar *et al.* (2019)]. In most cases, our proposed GSDP-GridNet model produces higher quality depth maps for generating 3D ken burns effect compared from other state-of-the-art methods. Apart from [Niklaus *et al.* (2019)] the rest of the papers do not highly focus on geometric, semantic and depth boundary distortion that is why they unable to synthesize perceptually good 3D ken burns effect.

In Fig 5.1 there are semantic distortions in [Niklaus *et al.* (2019)] depth where the depth of deer head is not separated from the background. Other methods have many semantic and geometric distortions which results in objects being teared apart and stretched/expanded unnaturally in the 3D ken burns effect.

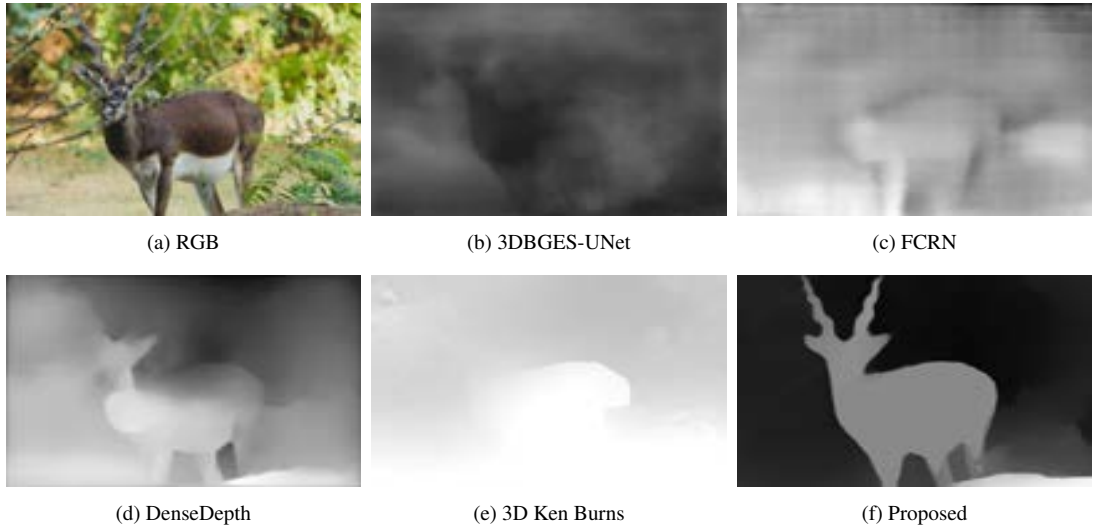


Figure 5.1: Comparison of Depth maps on Deer Scene.

In Fig 5.2 there is depth boundary error in depth maps of FCRN and DenseDepth. Ours and [Niklaus *et al.* (2019)] depth are equally good but the later produces better ken burn effect in this scene.

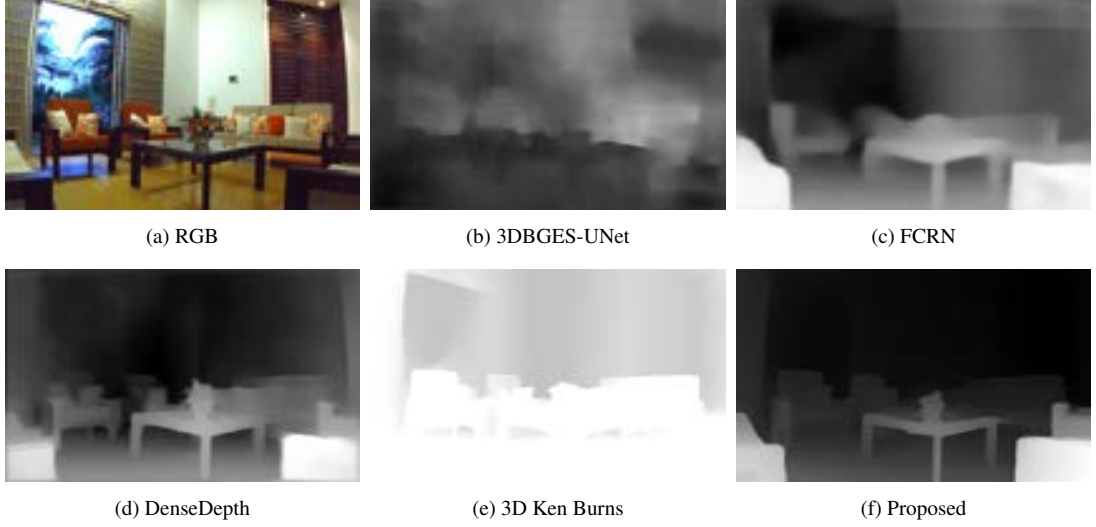


Figure 5.2: Comparison of Depth maps on Hall Scene.

In Fig 5.3 there is geometric distortion in FCRN and DenseDepth depth maps where the geometry of the left side is not visible and it directly affects the synthesis of ken burn effect. The [Niklaus *et al.* (2019)] depth has semantic distortion as the depth of the man is not consistent throughout the body. Ours depth map shows clear separation of the man from the background.

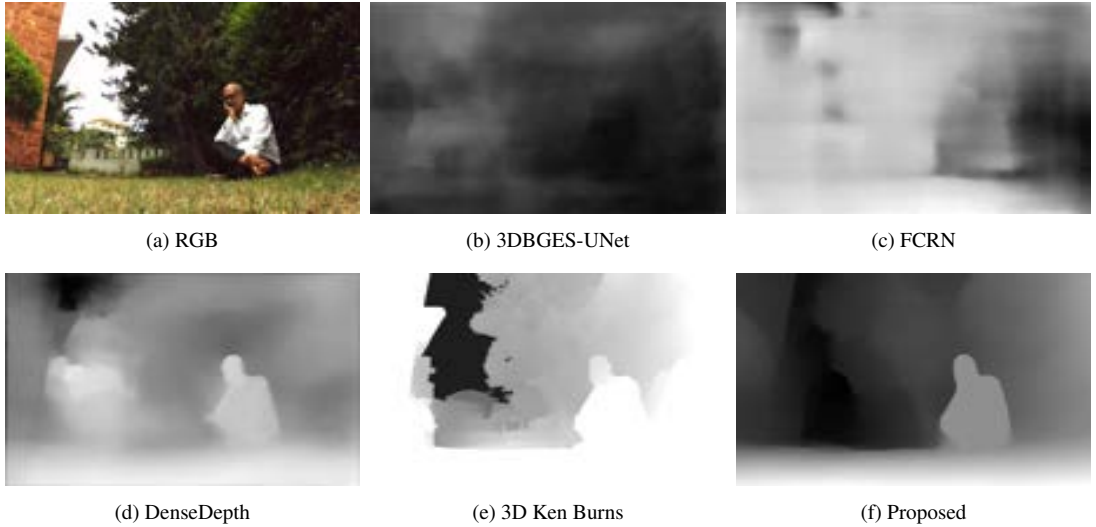


Figure 5.3: Comparison of Depth maps on Baba Scene.

Fig 5.4 has a complex depth structure with large depth variations, and high frequency regions of tree bushes. Apart from our depth map, others fail to capture the depth of canopy properly. Many semantic distortions can be observed in the depth maps of others which results in canopy being detached from the pole in the synthesized 3D ken burns effect.

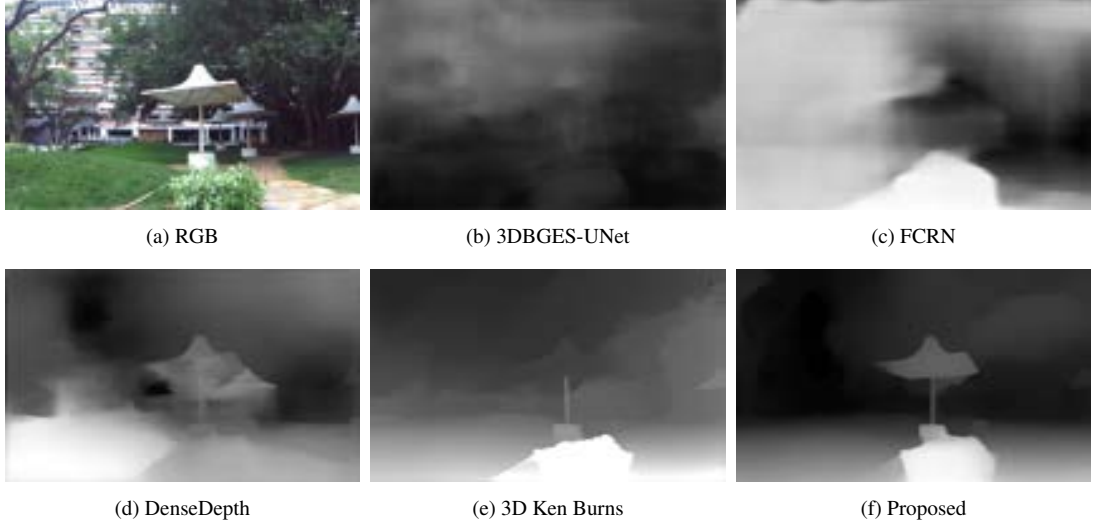


Figure 5.4: Comparison of Depth maps on NAC Canopy Scene.

In Fig 5.5 the [Niklaus *et al.* (2019)] depth map fails to capture the bush in front of the tree. Only our method gives depth map which captures the bush depth. Other methods fails to even capture the basic geometry of the scene which results in unnatural stretching of the tree branches in the synthesized 3D ken burns effect.

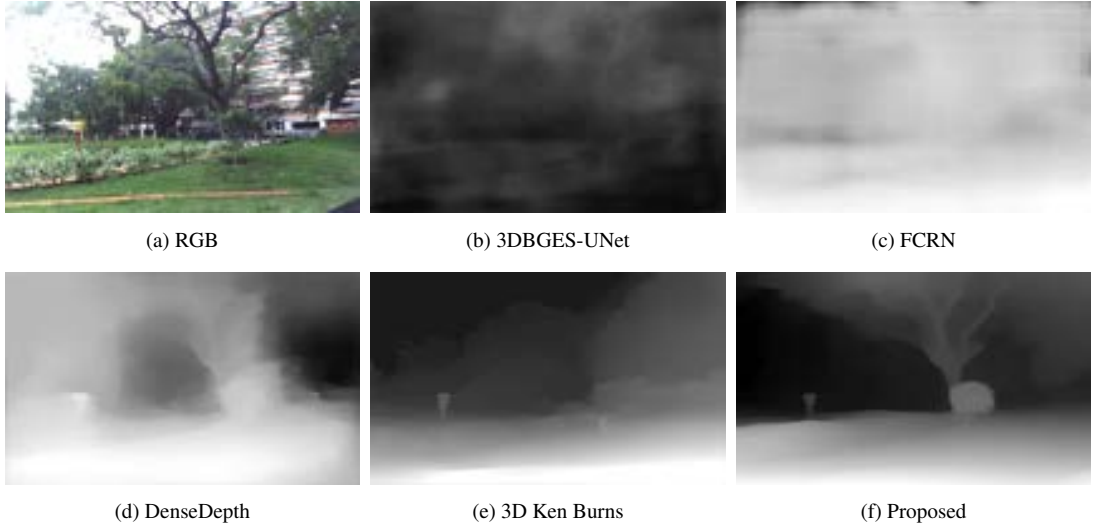


Figure 5.5: Comparison of Depth maps on NAC Trees Scene.

Fig 5.6 also has complex depth structure with large depth variation, minute details of trees and rain water reflections. Semantic distortion can be seen in [Niklaus *et al.* (2019)] depth map where the depth of bus top is not consistent. Depth boundary distortion can also be seen in FCRN and DenseDepth depth map which results in flickering issue in the synthesized ken burns effect.

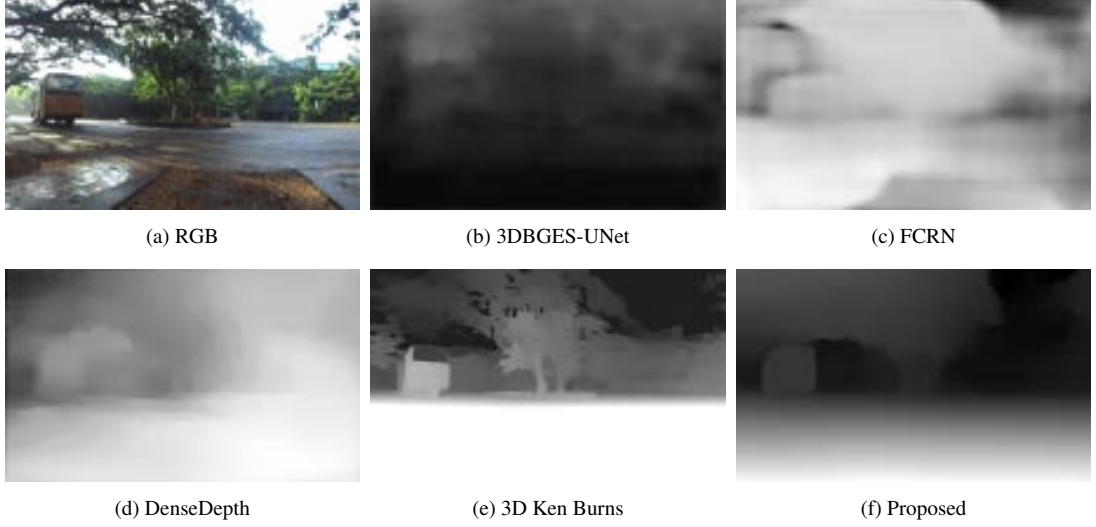


Figure 5.6: Comparison of Depth maps on Jamuna bus stand Scene.

5.2 View Synthesis Results

The view synthesis results obtained using the depth maps of 3DBGES-UNet[Mansi *et al.* (2020)], FCRN[Laina *et al.* (2016)], DenseDepth[Alhashim and Wonka (2018)], 3D Ken Burns [Niklaus *et al.* (2019)], Ours GSDP-GridNet method are shown in the below figures.

In Fig 5.7 the left vertical wall and window shed gets bent, the tree branches and bushes get unnaturally stretched in the ken burns effect while using 3DBGES-UNet, FCRN and DenseDepth depth maps. The result from 3D ken burns [Niklaus *et al.* (2019)] has the issue of blurring artifacts in the disocclusion areas.



Figure 5.7: Snapshots of 3D ken burns effect of Baba Scene.

In Fig 5.8 geometric artifacts appear where deer horns get blended with the background for 3DBGES-UNet depth. The body and neck get unnatural expansion for DenseDepth and FCRN depth. While using 3D ken burns depth we observe semantic distortions where the head of the deer is torn apart from the body in the synthesis of 3D ken burns effect because of inconsistent depth predictions in Fig. 5.1.

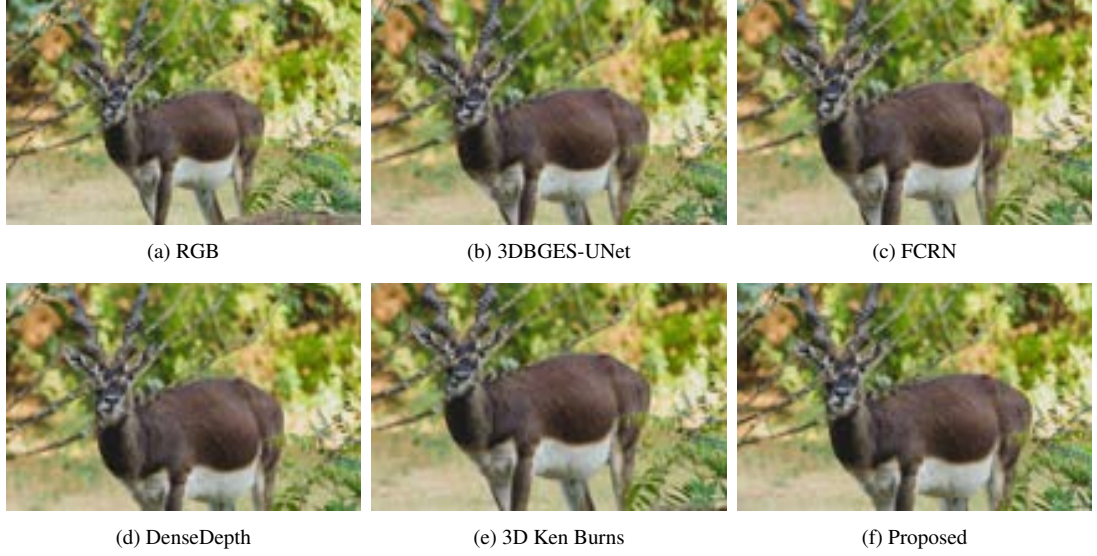


Figure 5.8: Snapshots of 3D ken burns effect of Deer Scene.

In Fig 5.9 geometric distortions of the bus and tree branches can be seen in the ken burns effect generated from depth maps of 3DBGES-UNet, DenseDepth and FCRN. Some flickering artifacts in the novel view renderings of 3D Ken Burns are also visible around the outer body of the bus due to inaccurate depth boundaries of the object.



Figure 5.9: Snapshots of 3D ken burns effect of Jamuna bus stand Scene.

Artifacts like stretching and expansion of objects occur due to incorrect depth orderings in predicted geometric relations. The rendering separates the parts of the object from the rest of the object because of inconsistent depth estimation inside the region of the object which sometimes make it indistinguishable from its background. That eventually results in part of the objects sticking to the background and being torn apart.

Our proposed method generates 3D ken burns effect that are pleasing to the eye with unnoticeable distortions compared to other methods. We also carried out an informal subjective user study to find out which method gives the best experience in Fig 4.1.

CHAPTER 6

Conclusion

In this work, we built a fully automated framework to generate a 3D ken burn effect from a single image in real-time. Our method consists of depth prediction pipeline which predicts geometric and semantic consistent depth map with accurate depth boundaries and view synthesis pipeline to render novel views to generate the 3D ken burns video. Our novel depth prediction network makes explicit use of the semantic and geometric information to predict depth maps suitable for view synthesis. We experimented with a wide variety of images and showed that our method generates realistic 3D ken burn video. Our subjective analysis showed that our proposed fully automated framework generates better results compared to the manual and existing methods of generating 3D ken burns effect.

REFERENCES

1. **Alhashim, I. and P. Wonka** (2018). High quality monocular depth estimation via transfer learning. *arXiv e-prints*, **abs/1812.11941**. URL <https://arxiv.org/abs/1812.11941>.
2. **Didyk, P., P. Sitthi-Amorn, W. T. Freeman, F. Durand, and W. Matusik** (2013). Joint view expansion and filtering for automultiscopic 3d displays. *ACM Transactions on Graphics (Proceedings SIGGRAPH Asia 2013, Hong Kong)*, **32**(6).
3. **(DIML), D. I. M. L.** (2016). Diml/cvl rgb-d dataset: 2m rgb-d images of natural indoor and outdoor scenes. <https://dimlrgbld.github.io/t>.
4. **Flynn, J., I. Neulander, J. Philbin, and N. Snavely**, Deep stereo: Learning to predict new views from the world’s imagery. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016.
5. **Fourure, D., R. Emonet, E. Fromont, D. Muselet, A. Tremeau, and C. Wolf** (2017). Residual conv-deconv grid network for semantic segmentation. *arXiv preprint arXiv:1707.07958*, 1–12.
6. **Hedman, P., S. Alsisan, R. Szeliski, and J. Kopf** (2017). Casual 3d photography. *ACM Trans. Graph.*, **36**(6). ISSN 0730-0301. URL <https://doi.org/10.1145/3130800.3130828>.
7. **Hedman, P. and J. Kopf** (2018). Instant 3d photography. *ACM Trans. Graph.*, **37**(4). ISSN 0730-0301. URL <https://doi.org/10.1145/3197517.3201384>.
8. **Horry, Y., K.-I. Anjyo, and K. Arai**, Tour into the picture: Using a spidery mesh interface to make animation from a single image. In *Proceedings of the 24th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH ’97*. ACM Press/Addison-Wesley Publishing Co., USA, 1997. ISBN 0897918967. URL <https://doi.org/10.1145/258734.258854>.

9. **Klose, F., O. Wang, J.-C. Bazin, M. Magnor, and A. Sorkine-Hornung** (2015). Sampling based scene-space video processing. *ACM Trans. Graph.*, **34**(4). ISSN 0730-0301. URL <https://doi.org/10.1145/2766920>.
10. **Koch, T., L. Liebel, F. Fraundorfer, and M. Körner**, Evaluation of cnn-based single-image depth estimation methods. In **L. Leal-Taixé and S. Roth** (eds.), *European Conference on Computer Vision Workshop (ECCV-WS)*. Springer International Publishing, 2018.
11. **Laina, I., C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab**, Deeper depth prediction with fully convolutional residual networks. In *3D Vision (3DV), 2016 Fourth International Conference on*. IEEE, 2016.
12. **Li, Z. and N. Snavely**, Megadepth: Learning single-view depth prediction from internet photos. In *Computer Vision and Pattern Recognition (CVPR)*. 2018.
13. **Liu, G., F. A. Reda, K. J. Shih, T.-C. Wang, A. Tao, and B. Catanzaro** (2018a). Image inpainting for irregular holes using partial convolutions.
14. **Liu, M., X. He, and M. Salzmann**, Geometry-aware deep network for single-image novel view synthesis. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2018b.
15. **Mansi, S., A. Sharma, K. R. Tushar, and A. Pannier** (2020). A novel 3d-unet deep learning framework based on high-dimensional bilateral grid for edge consistent single image depth estimation. *arXiv preprint arXiv:2105.10129*, 1–8.
16. **Nathan Silberman, P. K., Derek Hoiem and R. Fergus**, Indoor segmentation and support inference from rgb-d images. In *ECCV*. 2012.
17. **Niklaus, S., L. Mai, J. Yang, and F. Liu** (2019). 3d ken burns effect from a single image. *ACM Transactions on Graphics*, **38**(6), 184:1–184:15.
18. **Prateek, S. and S. Mansi** (2021). 3D Ken Burns Effect Results. <https://sites.google.com/site/mansisharmaitd/publications/gsdg-gridnet>.
19. **Ramamonjisoa, M. and V. Lepetit** (2019). Sharpnet: Fast and accurate recovery of occluding contours in monocular depth estimation. *The IEEE International Conference on Computer Vision (ICCV) Workshops*.

20. **Simonyan, K. and A. Zisserman** (2014). Very deep convolutional networks for large-scale image recognition. *CoRR*. URL <http://arxiv.org/abs/1409.1556>.
21. **Srinivasan, P. P., T. Wang, A. Sreelal, R. Ramamoorthi, and R. Ng**, Learning to synthesize a 4d rgb-d light field from a single image. *In 2017 IEEE International Conference on Computer Vision (ICCV)*. 2017.
22. **Tulsiani, S., R. Tucker, and N. Snavely**, Layer-structured 3d scene inference via view synthesis. *In Proceedings of the European Conference on Computer Vision (ECCV)*. 2018.
23. **Vladimir Nekrasov, I. R., Chunhua Shen**, Light-weight refinenet for real-time semantic segmentation. *In CVPR*. 2018. URL <https://arxiv.org/abs/1810.03272>.
24. **Wadaskar, A., S. Mansi, and R. Lal** (2019). A rich stereoscopic 3d high dynamic range image video database of natural scenes.
25. **Xie, J., L. Xu, and E. Chen**, Image denoising and inpainting with deep neural networks. *In F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger* (eds.), *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012. URL <https://proceedings.neurips.cc/paper/2012/file/6cdd60ea0045eb7a6ec44c54d29ed402-Paper.pdf>.
26. **Yeh, R. A., C. Chen, T. Y. Lim, A. G. Schwing, M. Hasegawa-Johnson, and M. N. Do**, Semantic image inpainting with deep generative models. *In 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.
27. **Zheng, K., A. Colburn, A. Agarwala, M. Agrawala, D. Salesin, B. Curless, and M. Cohen**, Parallax photography: Creating 3d cinematic effects from stills. volume 2009. 2009.