

Viewpoint Recommendation for Aesthetic Photography

by

Sathish Kumar Katukuri

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Science

Approved November 2019 by the
Graduate Supervisory Committee:

Robert LiKamWa, Chair
Pavan Turaga
Suren Jayasuriya

ARIZONA STATE UNIVERSITY

December 2019

ABSTRACT

This thesis addresses the problem of recommending a viewpoint for aesthetic photography. Viewpoint recommendation is suggesting the best camera pose to capture a visually pleasing photograph of the subject of interest by using any end-user device such as drone, mobile robot or smartphone. Solving this problem enables to capture visually pleasing photographs autonomously in areal photography, wildlife photography, landscape photography or in personal photography.

The viewpoint recommendation problem can be divided into two stages: (a) generating a set of dense novel views based on the basis views captured about the subject. The dense novel views are useful to better understand the scene and to know how the subject looks from different viewpoints and (b) each novel is scored based on how aesthetically good it is. The viewpoint with the greatest aesthetic score is recommended for capturing a visually pleasing photograph.

DEDICATION

Dedicated to my parents and sisters

ACKNOWLEDGMENTS

Firstly, I thank my parents and sisters for supporting and encouraging me to reach this wonderful place and pursue my masters degree.

I greatly appreciate the opportunity provided by Prof. Robert LiKamWa. The valuable inputs and feedback he provided helped me grow as a person both professionally and personally. The suggestions he provided helped me to fine tune my problem solving approach in the right direction.

A special mention to all my lab seniors Venkatesh Kodukula, Jinhua Hu, and Frank Liu for their warm suggestions and advice.

TABLE OF CONTENTS

	Page
LIST OF TABLES	vi
LIST OF FIGURES	vii
CHAPTER	
1 INTRODUCTION	1
2 RELATED WORK	5
2.1 View Synthesis	5
2.2 Aesthetics Evaluation	7
3 PROBLEM DESCRIPTION	8
4 METHODOLOGY	9
4.1 VIEW SYNTHESIS	10
4.1.1 Collecting Basis Views	10
4.1.2 3D Pose Extraction	11
4.1.3 MPIs Generation	13
4.1.4 Novel Views Generation	13
4.1.5 Render Novel Views	15
4.2 AESTHETICS EVALUATION	16
5 EXPERIMENTS AND RESULTS	19
5.1 Experimental Setup	19
5.2 Performance Evaluation	19
5.3 User Survey Design	20
5.4 User Survey Results	22
5.5 Results Without View Synthesis	30
6 CONCLUSIONS AND FUTURE WORK	32
6.1 Conclusions	32

CHAPTER	Page
6.2 Limitations and Future Work	32
REFERENCES	33

LIST OF TABLES

Table	Page
5.1 User Survey Results	22
5.2 Image Attribute Scores for Prediction Failure - 1	28
5.3 Image Attribute Scores for Prediction Failure - 2	29
5.4 Basis View and Novel View Scores	31

LIST OF FIGURES

Figure	Page
1.1 Bad Viewpoint and Good Viewpoint	2
1.2 Golden Spiral	3
2.1 Multi Plane Image	6
4.1 Complete Block Diagram	9
4.2 Basis Views	10
4.3 Novel Views	11
4.4 Scene	12
4.5 MPIs at Different Depths	14
4.6 Render Novel Views	15
4.7 Heat-map	18
5.1 Natural Scenes	20
5.2 Non-Natural Scenes	21
5.3 Predicted Rating and User Rating Correlation	27
5.4 Captured Viewpoint with Smartphone Camera	30
5.5 Basis View and Novel View Comparison	31

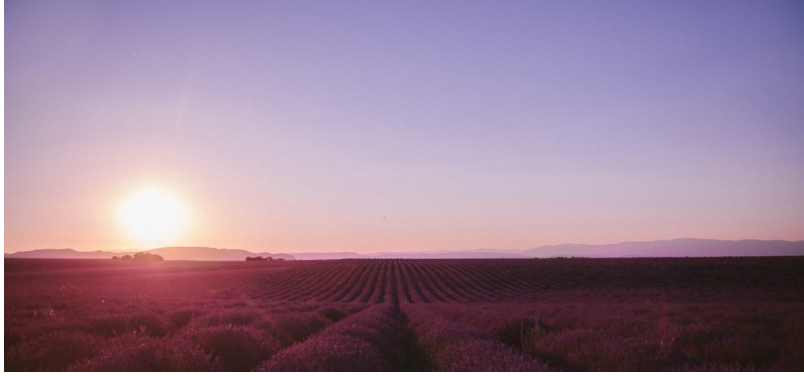
Chapter 1

INTRODUCTION

People like to take good photographs to capture moments in their life. To capture a visually pleasing photograph a better viewpoint is also required along with a good camera and good composition. Consider the images in Figure 1.1, both images are captured with the same camera and the same composition but with a different viewpoint. They are both good photos. The first one shown in Figure 1.1a seems to focus on the sunset, and the lavender is just there. The second shot shown in Figure 1.1b highlights the lavender rows. By capturing photographs from a better viewpoint we can produce more compelling photographs of a scene but for a typical camera user, it is difficult to know a better viewpoint. In this thesis, a framework is proposed to recommend a viewpoint for any end-user device (drone, ground robot or smartphone) that could be navigated to capture a visually pleasing photograph.

A closely related problem, robotic photography which deals with capturing well-composed photographs using mobile robots [1], [2], [3] and [4] use face detection to identify the human subjects. Once the subject is identified the input or basis view is collected and a set of views (candidate views) near the basis view are generated. The aesthetic score of the candidate views are evaluated using a subset of following well known image composition rules [5].

- Rule of Thirds: If you place the points of interest at the intersection or along the lines which divides the image into nine equal parts, the photo becomes more balanced [6].



(a) Bad Viewpoint



(b) Good Viewpoint

Figure 1.1: Bad Viewpoint and Good Viewpoint

- Golden Spiral: Once we start splitting the image into rectangles by the golden ratio (approximately 1.618 to 1) forever we get the rectangles as shown in Figure 1.2. The subject should be placed on the smallest rectangle for better composition [7].
- Visual Balance: Framing the visually salient features such that objects and colors have equal visual weight creates a more balancing image [7].
- No-Middle: When the subject is placed right or left of the frame more balancing images can be created using background and other objects in the scene [7].

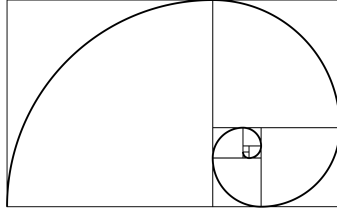


Figure 1.2: Golden Spiral

- Empty Space: Should not be more than two-thirds of empty space on the image [7].
- No Edges: Subject should not be at the edges of the frame [7].

The robot navigates to the candidate view with the greatest aesthetic score and captures the image. In these methods, aesthetics are evaluated for the candidate views which are generated based on a single basis view. As a single basis view is collected it does not have much visual information about the scene, so the candidate views generated based on this view are restricted to smaller viewing space. The systems are restricted to choose the best view from the smaller viewing space, therefore the systems are easily trapped in local maxima. In our approach, multiple basis views are collected around the subject and dense candidate views are generated on larger viewing space (180° around the subject) which helps to reach global maxima.

The content of an image is an important factor in deciding which attribute is more relevant for that image. For example, the Rule of Thirds and Golden Spiral are highly relevant in landscape images rather than closeup portraits. Neural networks are better at learning from patterns in data, so these can be used in assigning weights to attributes based on the image content (pattern). Instead of using the handcrafted image composition rules for aesthetic score evaluation, a Deep Neural Network based approach [8] is used which assigns weights to attributes based on the image content.

In summary, a modular framework is presented for viewpoint recommendation using the integration of well-studied modules: view synthesis [9] and Convolutional Neural Network(CNN) based aesthetic score evaluator [8]. The view synthesis module takes 30-40 basis views around the subject and generates multi-plane images(MPIs) [10], which are used to render dense novel views around the subject. The rendered novel views' aesthetic scores are calculated using the score evaluator. The 6DoF pose of rendered view with the greatest score is recommended for capturing the photograph by the end-user device.

The organization of the thesis is as follows.

- Chapter2 discusses related topics
- Chapter3 presents problem description
- Chapter4 describes the methodology
- Chapter5 describes the experiments conducted and results
- Chapter6 conclusions and future work

Chapter 2

RELATED WORK

The entire framework is divided into two stages, view synthesis and aesthetics evaluation. Existing MPI based CNNs are used for view synthesis [11] and we describe an algorithm to render dense novel views on a grid surface around the subject. In the second stage, an attribute based deep neural network [8] is used for aesthetics evaluation.

2.1 View Synthesis

View synthesis is the process of rendering novel views from different camera viewpoints by processing a set of basis views. Seitz et al [9] proposed a view synthesis method that uses three steps (1) 3D reconstruction from the basis views, (2) apply 3D scene, camera and illumination transformations, and (3) rendering novel views. However, this approach generates novel views only on the straight line connecting basis views. Other classical image transformation approaches [12], [13] also uses 3D reconstruction of the scene to generate novel views. While all these image-based methods yield high-quality novel views, they lack a systematic procedure to collect basis views and also limited to a smaller extrapolation of novel views.

The more recent learning-based approaches use the powerful deep learning framework to solve view synthesis. DeepSterio [14] and Light Field Synthesis [15] proposed CNNs to render novel views. However, these approaches predict each novel view separately by using only the basis views which results in geometrical inconsistencies in novel views. Stereo magnification [10] uses multiplane images(MPIs) as shown in Figure 2.1 to render novel views. MPIs are multi level planar representation of

fixed depth scene, the scene divided into fixed number of planes (32 or 64 or 128) each plane represent as RGBA image for example $(C_1, \alpha_1) \dots (C_D, \alpha_D)$, where C_d is the RGB image at depth d , α_d is the alpha/transparency at depth d and D is the maximum number of planes. To render the novel views they use planar transformation that inverse wraps each MPI RGBA layered representation to target view point. While this method generates high-quality, geometrically consistent rendered views, it is restricted to extrapolate the novel views in smaller base-line view from stereo pair and restricted to 1D camera path.

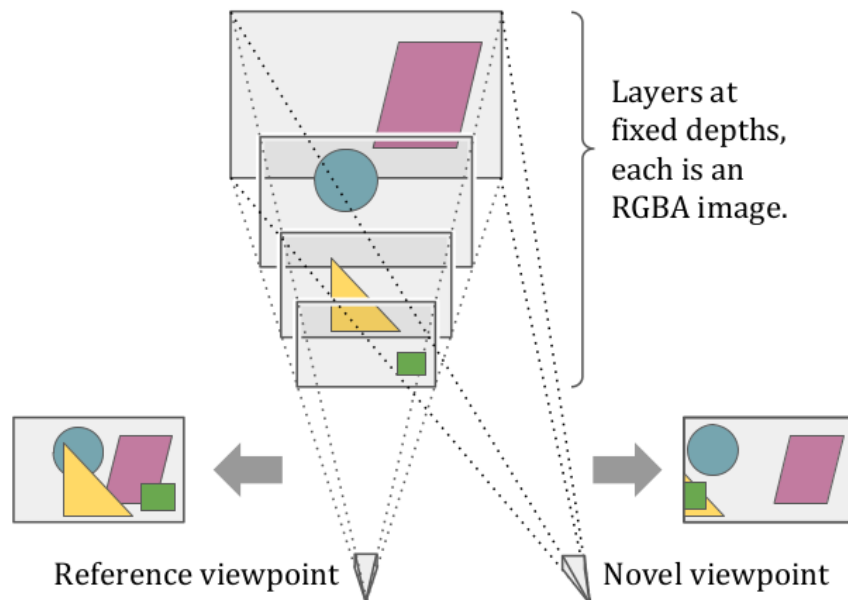


Figure 2.1: Multi Plane Image [10]

Local Light Field Fusion [11] is a 3D CNN improved up on MPI based [10] view synthesis for a larger base-line view from multiple basis views. This method supports dynamic adjustment of depth planes based on input view sampling rate which reduces the number of basis views required for high-quality rendering. The proposed approach in this method, to collect the basis views on a grid-like surface helps to avoid the

trial and error approach in collecting the basis views. The ability to generate novel views on 2D camera paths enables us to generate dense novel views on a 2D grid or semi-cylindrical surface around the subject for better scene understanding. In this work, this method is adopted and custom camera trajectory algorithms are used for rendering dense novel views around the subject.

2.2 Aesthetics Evaluation

Image aesthetic evaluation is a subjective activity, human judgment can be affected by personal taste. However, there are some widely accepted computational aesthetics [3] such as Rule of Thirds and Golden Ratio. The computational aesthetics of an input image can be improved by changing the relative position of salient features using mathematical operators like crop-and-retarget.

The subjective aesthetics such as *interesting content* in an image lack clear definition, so it is difficult to implement them computationally. A considerable amount of research has been done on the evaluation of subjective aesthetics. Aesthetic Visual Analysis (AVA) [16] released a dataset with synthetic and natural images and evaluated the aesthetics based on attributes *interesting content*, *object emphasis*, *good lighting*, *color harmony*, *vivid color*, *shallow depth of field*, *motion blur*, *rule of thirds*, *balancing element*, *repetition*, and *symmetry*. However, AVA attributes are binary, which means it can only tell whether a particular attribute is present or not in an image. Photo Aesthetics Ranking Network with Attributes and Content Adaptation(AADB) [8] evaluates AVA attributes on a scale of -1 to 1, which helps to understand which attribute is affecting the aesthetics of an image. In this work, AADB CNN is used to evaluate the aesthetic score of a viewpoint.

Chapter 3

PROBLEM DESCRIPTION

The basis views (30-40) are collected about the subject in a grid-like surface and a set of dense novel views are generated around the subject using the basis views. For each basis view, an MPI is generated. The images for each novel view are rendered by wrapping four nearest MPIs of the target novel view. The rendered novel views are evaluated for aesthetics and the viewpoint with the greatest score is recommended for capturing the image. The aforementioned solution developed with the following assumptions.

1. There should be significant visual overlap in the basis views, i.e every object of the scene should be visible in at least 3 basis views. The basis views should be captured from different viewpoints.
2. The basis views should include the views on the boundaries of the scene for an accurate representation of the scene from different viewpoints.
3. Basis views should be captured in similar lighting conditions.
4. The maximum disparity between basis views should not exceed 64 pixels.

METHODOLOGY

The entire framework consists of two stages, View Synthesis and Aesthetics Evaluator as shown in Figure 4.1. The collected basis views are given as input to 3D Pose Extractor and MPI Generator. The 3D Pose Extractor extracts the camera pose and the MPI Generator generates MPI for each basis views. The extracted camera poses are used by the Generate Novel View module to generate dense novel views on the grid-like surface. Render Novel Views module uses the MPIs and dense novel views to render images. The rendered images aesthetic scores are evaluated by Aesthetic Evaluator and the pose of the viewpoint with the greatest aesthetic score is returned.

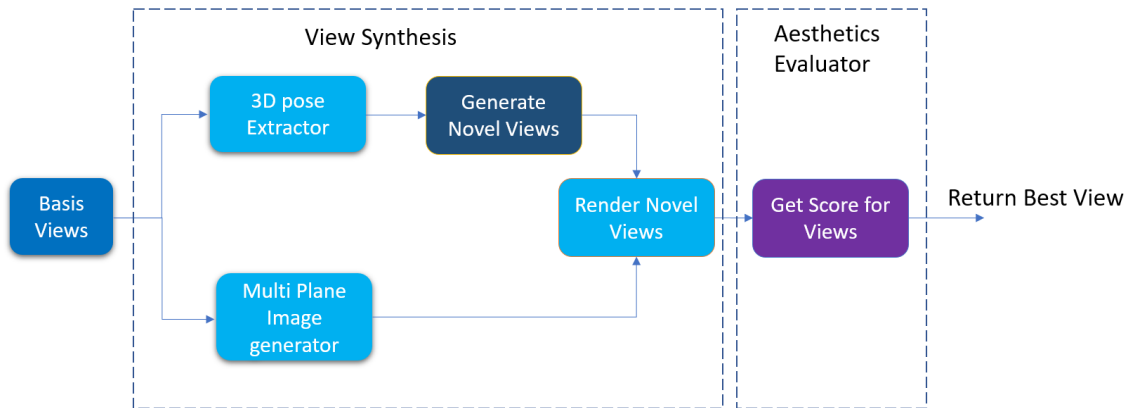


Figure 4.1: Complete Framework

4.1 VIEW SYNTHESIS

4.1.1 Collecting Basis Views

Collecting basis views is one of the crucial steps for 3D pose extraction, there should be significant overlap between two basis views at the same time they should not represent the same view. Mildenhall et al. [11] proposed a systematic procedure to collect the basis views on a grid-like pattern which satisfies both the conditions for 3D pose extraction. By collecting 30-40 basis views around the subject on a grid pattern as shown in Figure 4.2, the novel views are generated as shown in Figure

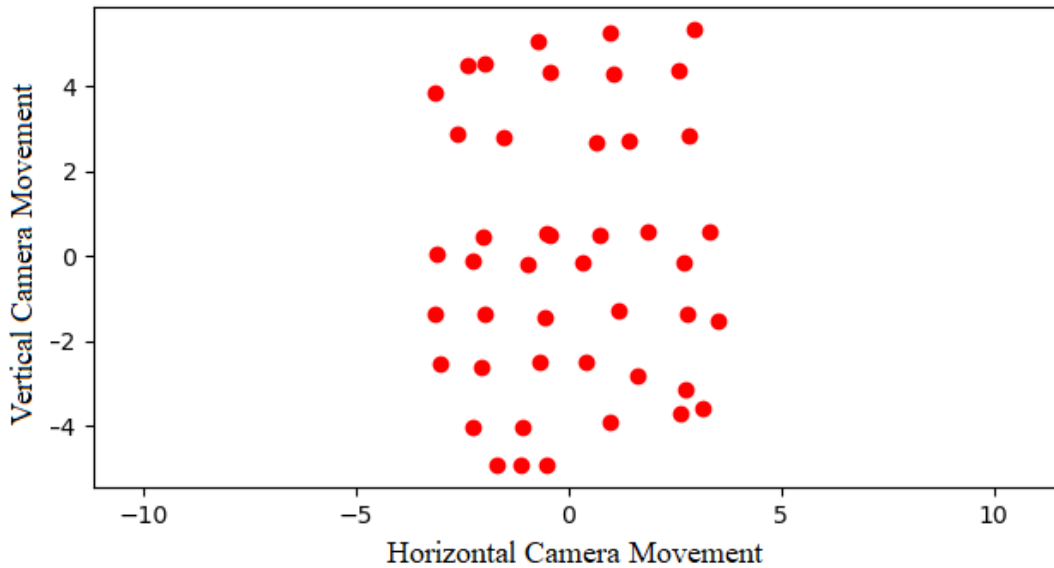


Figure 4.2: Basis Views

4.3 for the given scene Figure 5.3. The number of basis views required is calculated using the Equation 4.1 [11], where W is the target render image width, N is the number of basis views required, z_{min} closest scene depth and S is the side length of

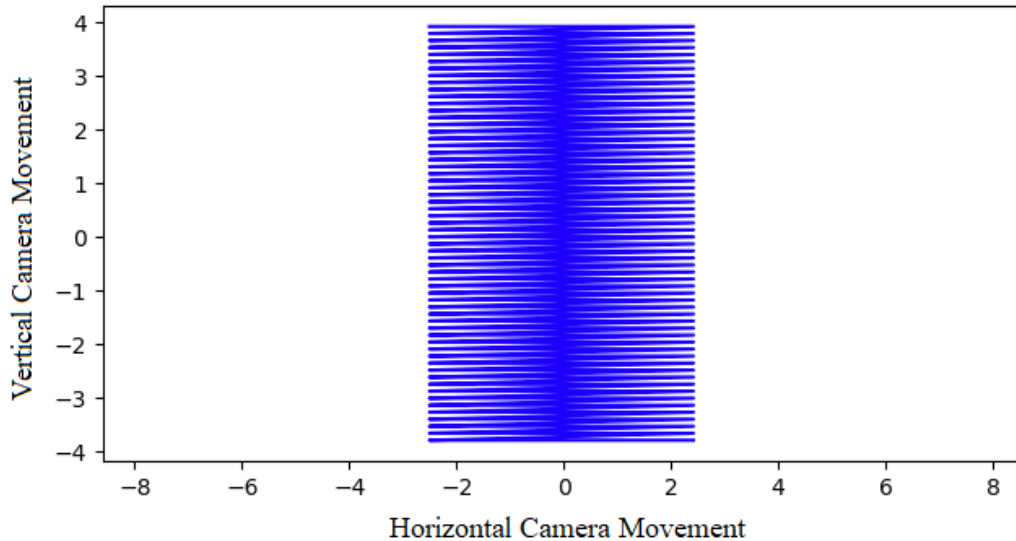


Figure 4.3: Novel Views

the view-space user wishes to render. For a closest scene depth of 2 meters, target render image width of 480 pixels and the view-space width of 2 meters the number of basis views required is ≤ 36

$$W/\sqrt{N} \leq 80z_{min}/S \quad (4.1)$$

4.1.2 3D Pose Extraction

The COLMAP [17] library is used to extract the 6DoF camera pose in the world for each basis view. COLMAP is an image-based 3D reconstruction library which recovers the sparse reconstruction of the scene and the camera poses of the input images using Structure-from-Motion(SfM) [18]. The input is a set of overlapping images of the same scene from different viewpoints. The output is a 3D reconstruction of the scene and the reconstructed intrinsic, extrinsic camera parameters of all the



Figure 4.4: Scene

images. The following are the steps involved in extracting the pose.

- The input images are collected as described in the Section 4.1.1
- The camera intrinsics are extracted from the image Exchangeable Image File Format(EXIF) information. If an image has partial EXIF information COLMAP automatically finds the missing camera parameters using simple radial distortion model (simplified version of the OPENCV [19] model only modeling radial distortion effects with one parameter). The parameters are refined during the sparse reconstruction.
- The SIFT [20] features are detected and extracted from the input images.

- Exhaustive matching is done to find the correspondences between the feature points in different images.
- The camera poses are extracted while doing the sparse reconstruction by triangulating the feature points.

On an NVIDIA Tesla V100, it takes about 4-6 minutes to process the basis views and extract the poses

4.1.3 MPIs Generation

Using the collected basis views as input to the pre-trained CNNs [11], the MPIs are generated for each basis views. To generate MPI for a reference view, 4 nearest neighbors along with the reference view are re-projected on to D (32, 64 or 128) planes to form 5 volumes of each $H \times W \times D \times 3$. The CNN takes these volumes as the input and generates a set of 5 color selections weights and opacity α for each MPI coordinate (x, y, α) . The weights are used to calculate each MPI coordinate RGB color value as a weighted combination of 5 input image coordinates color value. Figure 4.5 shows the generated MPIs (an $RGB\alpha$ image at depth d) at different depth of the scene Figure 5.3. On an NVIDIA Tesla V100, in total it takes 5-7 minutes to generate MPIs for a 30-40 basis view with 360x480 and 32 planes resolution.

4.1.4 Novel Views Generation

The best viewpoint to take a visually pleasing photograph of the scene can only be decided if we know how the scene looks form different viewpoints. The novel views as shown in Figure 4.3 for the scene shown in Figure 5.3 were generated using below algorithm.

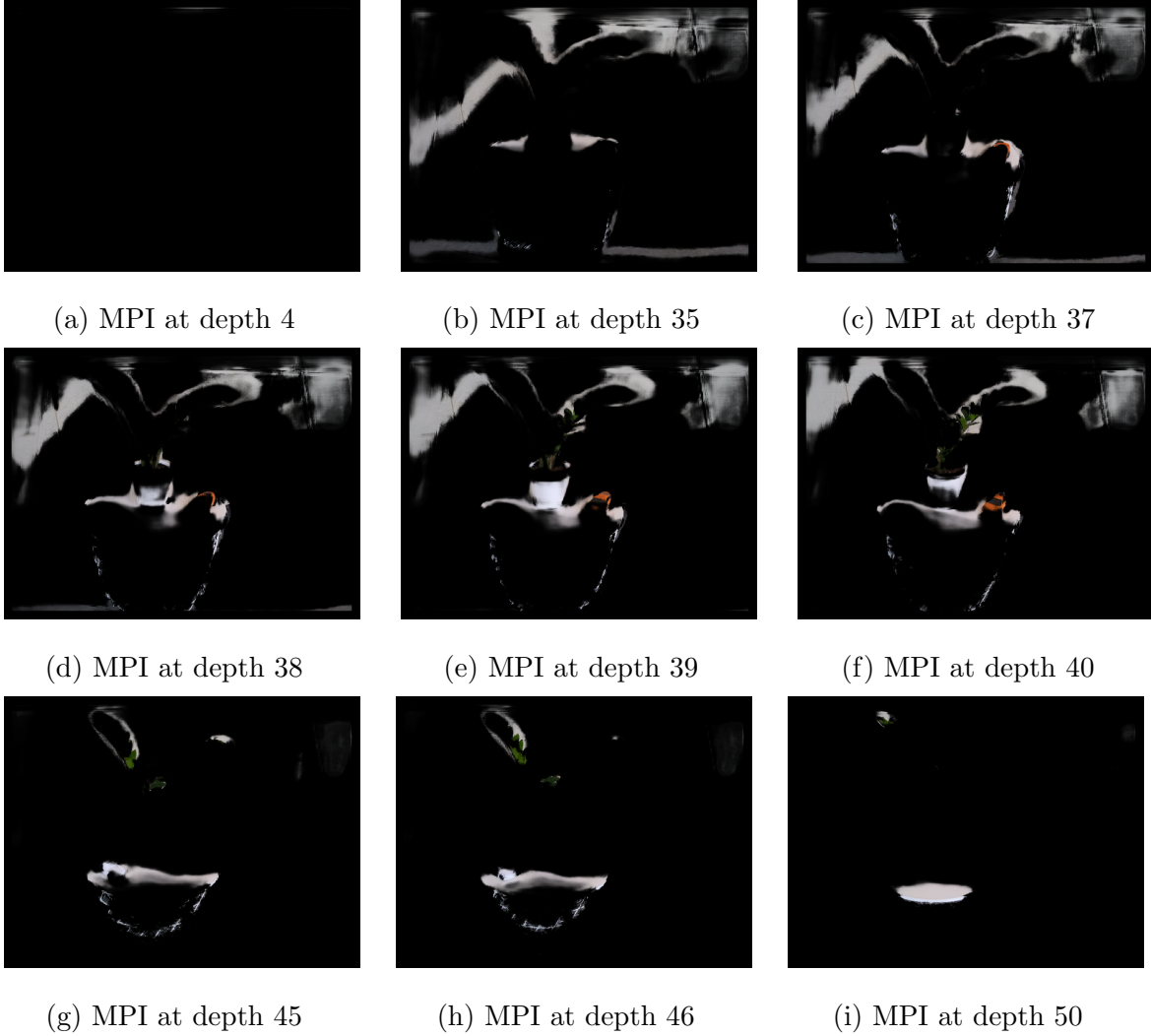


Figure 4.5: MPIs at Different Depths

- Get the average of all the basis views camera poses i.e C2W (camera to world transformation) which is nothing but camera center in world coordinate system.
- Generate the novel views on 2D space, x-axis being horizontal movement and the y-axis being the vertical movement around the camera center $(0, 0, 0)$ in the camera coordinate system.

- Transform the generated novel views in the above step to the world coordinate system by applying the camera to world transformation (C2W).
- Collect all the poses from the above step to render images for the novel views.

4.1.5 Render Novel Views

To render a target novel view at pose p_t using the MPIs generated in Section 4.1.3, each $\text{RGB}\alpha$ plane is wrapped onto the target pose frame and the alpha composition is done from back to front. The planar transformation that maps the basis MPI $\text{RGB}\alpha$ onto the target viewpoint described by the equation 4.2 [10], where u_t and v_t are target image points u_s and v_s are source basis view image points, the 3D transformation matrix from source to target is define by Rotation matrix R , translation t , the camera intrinsics are given by k_s and k_t and n denote the MPI plane normal. A single MPI alone will not contain all the visual information required for rendering the target view due to occlusion and field of view issues. The final RGB image at the target viewpoint is generated by blending multiple MPIs as shown in Figure 4.6

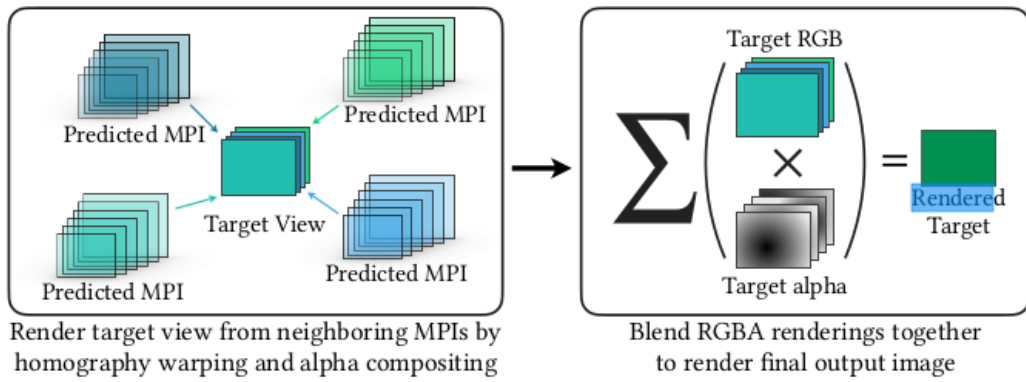


Figure 4.6: Render Novel Views [11]

$$\begin{bmatrix} u_s \\ v_s \\ 1 \end{bmatrix} \sim k_s \left(R^T + \frac{R^T t n R^T}{a - n R^T n} \right) k_t^{-1} \begin{bmatrix} u_t \\ v_t \\ 1 \end{bmatrix} \quad (4.2)$$

4.2 AESTHETICS EVALUATION

The rendered novel views in Section 4.1.5 are evaluated using the image aesthetic evaluator CNN [8]. The scores are evaluated for following aesthetic attributes [11].

- Content: How well the content is describing the emotion (or story) of the subject (or the scene).
- Object Emphasis: How well the image emphasizes foreground objects.
- Lighting: Whether the image has good/interesting lighting.
- Color Harmony: How well the colors that go together are used to create a pleasing image.
- Vivid Color: Whether the photo has vivid color, not necessarily harmonious color.
- Depth of Field: Depth of field is the distance between the closest and farthest objects in an image, both of which are in focus. Images with shallow depth of field provide more emphasis on the subject.
- Motion Blur: In photography, motion blur is the purposeful streaking or blurring of an object in motion for visual effect. The weight of this attribute is decided by the content of the image if the content is static this will have a negative effect else it will have a positive effect on the score.

- Rule of Third: If you place the points of interest at the intersection or along the lines which divides the images into nine equal parts, the photo becomes more balanced.
- Balancing Element: Framing the visually salient features such that objects and colors have equal visual weight creates a more balancing image.
- Repetition: When you repeat a certain size, shape or color you add strength and additional meaning to the overall image.
- Symmetry: Symmetry refers to a line that splits an object in half and, if both sides of the object are an exact mirror image of each other, then this object is said to be symmetrical. Symmetry lets you automatically create harmony and proportion in a photograph.

Each attribute is scored on the scale of -1 to 1, the negative score being the attribute has a negative effect, zero being no effect and the positive score being a positive effect on the image aesthetics. The embedded content-aware network assigns a weight to each attribute based on the content of the image. A final score is predicted as the weighted sum of attributes. The novel view with the greatest final prediction score selected as the best view and the corresponding 6DoF pose is recommended for capturing a visually pleasing photograph. Figure 4.7 shows heat-map based on the score for each view.

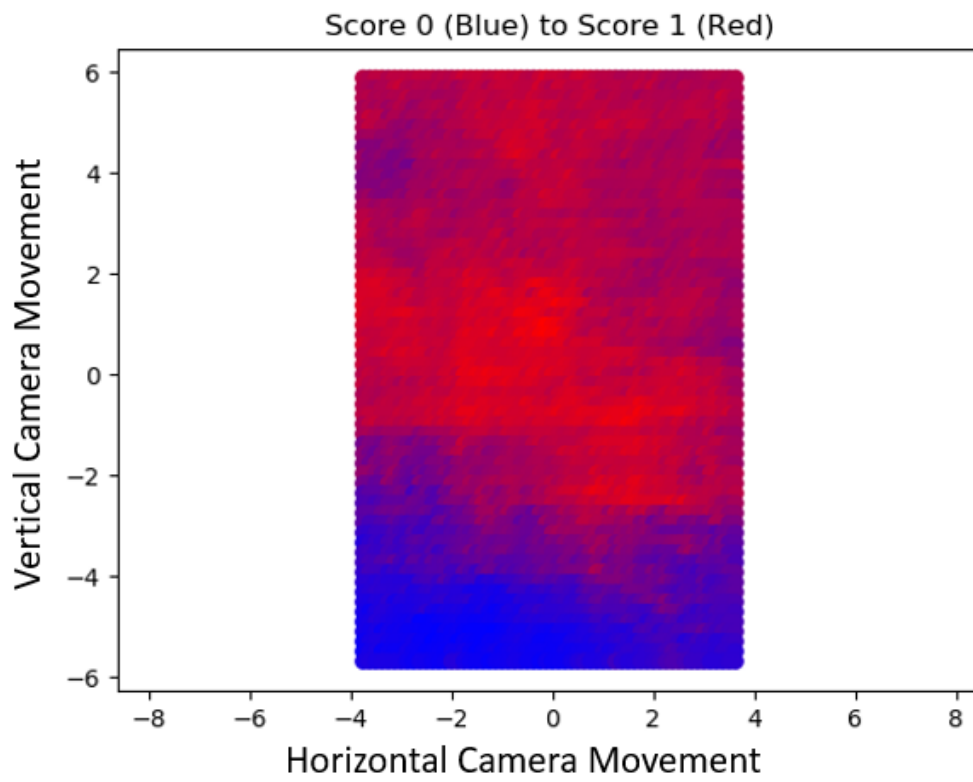


Figure 4.7: Heatmap

Chapter 5

EXPERIMENTS AND RESULTS

5.1 Experimental Setup

The experiments are designed to cover typical user case scenarios: Natural scenes (nature and landscape) as shown in Figure 5.1 and Non-Natural (man-made structure) scenes as shown in Figure 5.2. Using our pipeline, we collected basis views for each scene and generated novel views. Each novel view is scored based on its aesthetics using the aesthetic evaluator and we assign the predicted rating to the images based on the score given by the aesthetic evaluator on the scale of 1-10. To validate our predicted ratings we conducted a user survey with 31 users and also to know the degree of agreement between the user ratings and our predicted ratings we calculated the correlation coefficient as described in section 5.2. We developed an Android application to capture the recommended viewpoint by manually positioning the camera as described in section 5.3.

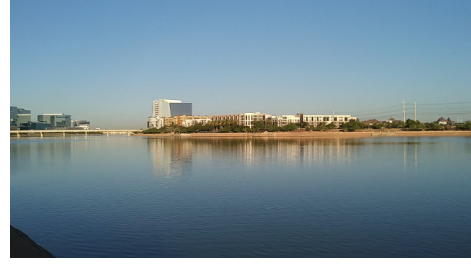
5.2 Performance Evaluation

As we are assessing the relationship between different rankings (predicted ratings by our pipeline and ratings given by the users in the survey) for the same input (images), the rank correlation measurement, Spearman Rank Correlation Coefficient (ρ) [21] is useful. It is a non-parametric correlation measurement which measures the strength and direction of the association between two rankings. The Spearman Rank Correlation Coefficient is described by equation 5.1, where

$$\rho = 1 - \frac{6 \sum d_i^2}{n^2(n-1)} \quad (5.1)$$



(a) 'A' Mountain



(b) Tempe Town Lake 1



(c) Cycle Track

Figure 5.1: Natural Scenes

- $d_i = r_i - \hat{r}_i$, r_i is the average rating given by the users to image i and \hat{r}_i is the predicted rating by our pipeline to image i .
- n is the number of users participated in the survey.

The Spearman correlation coefficient, (ρ), can take values from +1 to -1. A ρ of +1 indicates a perfect association of ranks, a ρ of zero indicates no association between ranks and a ρ of -1 indicates a perfect negative association of ranks. The closer ρ is to zero, the weaker the association between the ranks.

5.3 User Survey Design

The user survey is conducted on the natural scene shown in Figure 5.1b and the non-natural scenes shown in Figure 5.2c and Figure 5.2d. Users are provided with a total of 30 photographs with 10 photographs (rendered novel views) of each scene and asked to rate each photograph based on how much they like it on the scale of



(a) Gammage



(b) Horses



(c) Old Main



(d) Fountain



(e) Tempe Town Lake 2

Figure 5.2: Non-Natural Scenes

0-10. All the rendered novel views of each scene are sorted in descending order of scores and 10 images are selected for user survey as below

- 3 images randomly selected from the top 10 scored images
- 3 images randomly selected from bottom 10 scored images
- 4 images randomly selected from the remaining images

5.4 User Survey Results

Table 5.1 shows the images used in the user survey with their predicted rating and average user rating. In total 31 users participated in the survey.

Table 5.1: The table provides the user survey results participated by 31 users


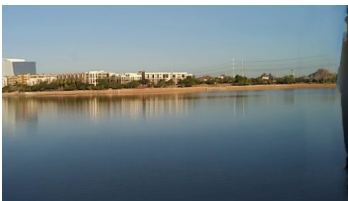
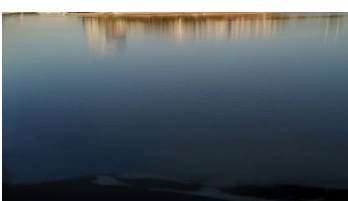
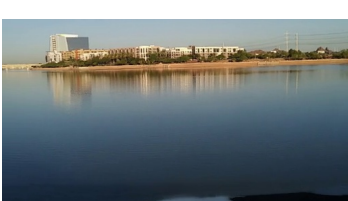
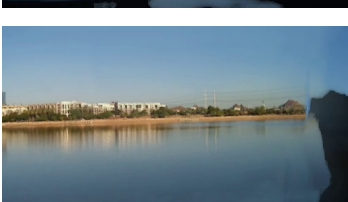
Image	Predicted Rating	User Rating
	8	7.2
	4	4.58
	1	2.83
	10	6.61
	6	3.09
Continued on next page		

Table 5.1 – continued from previous page


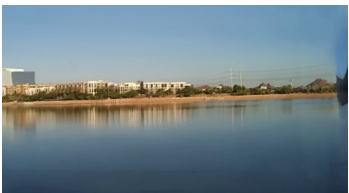

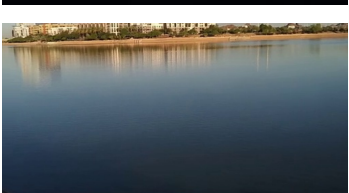
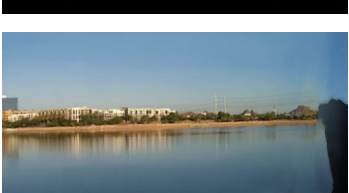
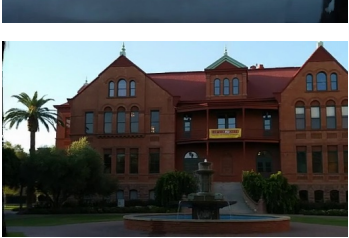
Image	Predicted Rating	User Rating (Average)
	8	6.51
	5	3.35
	1	1.96
	2	2.04
	4	2.87
	3	5.06
Continued on next page		

Table 5.1 – continued from previous page





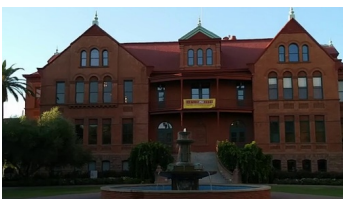

Image	Predicted Rating	User Rating (Average)
	10	4.32
	1	2.25
	2	4.0
	4	3.09
	5	6.06
	6	6.96
Continued on next page		

Table 5.1 – continued from previous page







Image	Predicted Rating	User Rating (Average)
	8	4.61
	9	3.32
	7	6.74
	9	7.41
	1	6.70
	2	6.90
Continued on next page		

Table 5.1 – continued from previous page








Image	Predicted Rating	User Rating (Average)
	4	6.45
	3	3.46
	5	5.58
	6	3.51
	7	4.93
	8	5.83
Continued on next page		

Table 5.1 – continued from previous page

Image	Predicted Rating	User Rating (Average)
	10	6.41

The Spearman Rank Correlation Coefficient (ρ) for the data presented in Table 5.1 is 0.732, this suggests that there is a strong correlation between the predicted aesthetics and the user evaluated aesthetics for a given viewpoint. Figure 5.3 shows the agreement trend between the predicted rating and the user rating (average). However, from the graph, there is a disagreement between the predicted rating and user rating for the images (Image-19 and Image-22) shown in Tables 5.2 and 5.3.

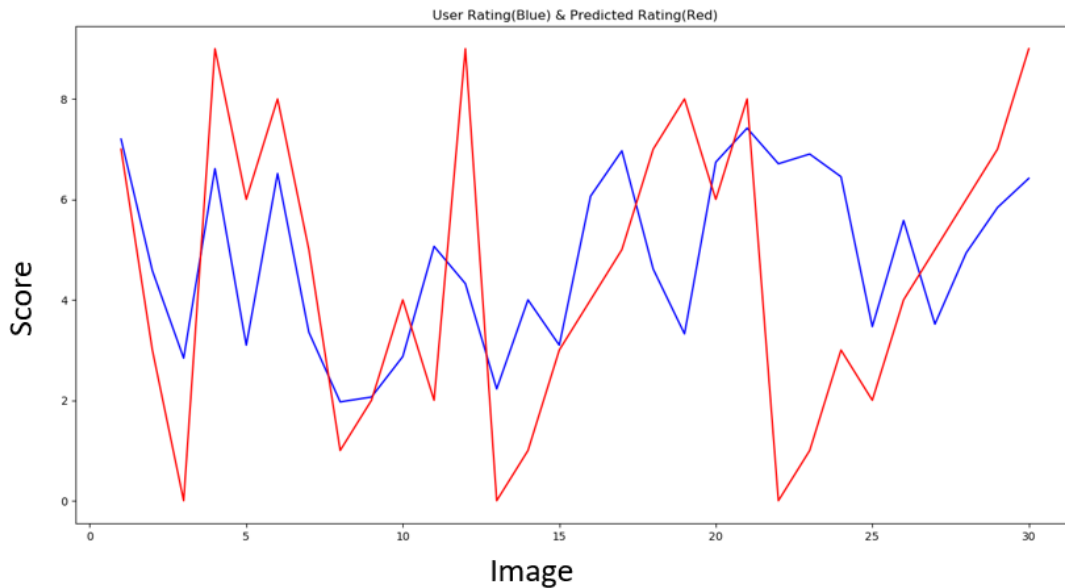


Figure 5.3: Predicted Rating and User Rating Correlation

- The image shown in Table 5.2 has been given the predicted rating of 1 but the user rating is 6.98. Even though the image is visually compelling the aesthetic evaluator assigned less score because the foreground and background objects are given equal visual weight, due to this image is penalized with a heavy negative score for the attribute *Object Emphasis*.


Image	Attribute Scores
	Final Predicted Score: 0.45114982,
	Color Harmony: 0.32407057,
	Content: -0.2810306,
	Vivid Color: 0.13903256,
	Repetition: 0.12058407,
	Shallow Depth of Field: 0.03755852,
	Light: -0.08843061,
	Motion Blur: -0.01240202,
	Balancing Element: 0.0205977,
	Object Emphasis: -0.6357091,
Symmetry: 0.09595232,	
Rule Of Thirds: 0.04955681	

Table 5.2: Image Attribute Scores for Prediction Failure - 1

- The image shown in Table 5.3 has been given the predicted rating of 9 but the user rating is 3. The aesthetic evaluator considered the black shadow part of the image as interesting content and assigned a heavy positive score for the attribute *Content*


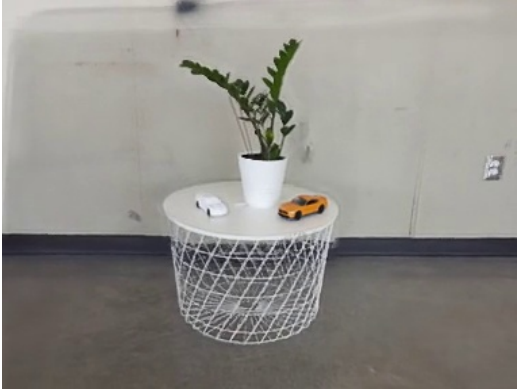
Image	Attribute Scores
	Final Predicted Score: 0.6400984,
	Color Harmony: 0.3136383,
	Content: 0.5788131,
	Vivid Color: 0.16890518,
	Repetition: 0.06972571,
	Shallow Depth of Field: -0.05330621,
	Light: 0.04557373,
	Motion Blur: -0.0249419,
	Balancing Element: -0.10086745,
	Object Emphasis: 0.24826978,
Symmetry: 0.20740245,	
Rule Of Thirds: 0.10703185	

Table 5.3: Image Attribute Scores for Prediction Failure - 2

To verify the recommended viewpoint, we developed an Android application to position the smartphone camera manually in the recommended viewpoint and capture the image. By following the on-screen feedback *to move in left, right, upwards, downwards, forward and backward directions*, we positioned the camera and captured the image. Figure 5.4a shows the recommended viewpoint by our framework, and Figure 5.4b shows the captured image by manually positioning the smartphone camera in the



(a) Recommended Viewpoint



(b) Captured Viewpoint

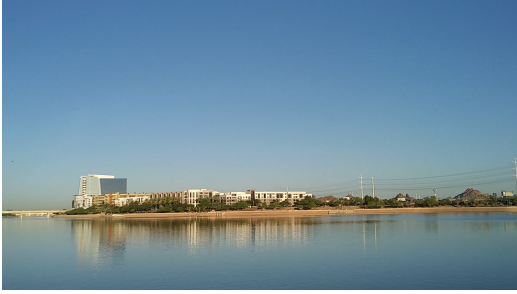
Figure 5.4: Captured Viewpoint with Smartphone Camera

recommended viewpoint. As we can see in Figures 5.4b and 5.4a the misalignment between the recommended view and captured images is due to the manual positioning of the camera. However, we believe that autonomous systems such as drones could capture the exact recommended viewpoint by positioning itself accurately.

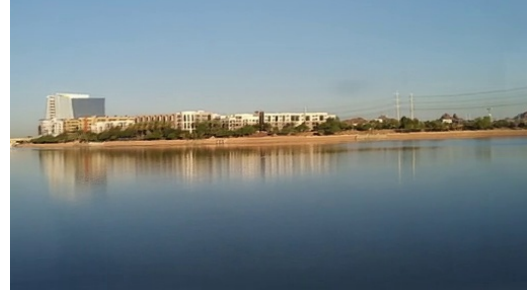
5.5 Results Without View Synthesis

The view synthesis stage in our framework takes about 10-15 minutes to render novel views using the basis views and the aesthetic evaluator stage works in real-time. To evaluate the trade-off between the quality of viewpoint and processing time we generated aesthetic scores for basis views (without view synthesis) and the novel views (with view synthesis). Figure 5.5 shows the highest scored basis view and novel view. The predicted scores for the basis view and novel view shown in Figure 5.5 are given in Table 5.4.

Form Table 5.4, the novel got a better final predicted score than the basis view. All the attribute scores are comparable between these two views except the *Rule of Thirds* attribute which is more relevant for landscape photography. The novel view



(a) Highest Scored Basis View



(b) Highest Scored Novel View

Figure 5.5: Basis View and Novel View Comparison

Basis View Score	Novel View Score
Final Predicted Score: 0.66290486	Final Predicted Score: 0.66499865
Color Harmony: 0.6689767	Color Harmony: 0.5098945
Content: 0.17461635	Content: 0.16503847
Vivid Color: 0.5019264	Vivid Color: 0.25542688
Repetition: 0.24444535	Repetition: 0.23503473
Shallow Depth of Field: -0.00712904	Shallow Depth of Field: 0.00095609
Light: 0.5319123	Light: 0.586401
Motion Blur: -0.15245223	Motion Blur: -0.07845964
Balancing Element: 0.29522672	Balancing Element: 0.34715205
Object Emphasis: -0.27588353	Object Emphasis: -0.12929033
Symmetry: 0.16660142	Symmetry: 0.22288111
Rule of Thirds: 0.00554072	Rule of Thirds: 0.26293015

Table 5.4: Basis View and Novel View Scores

got the better score for *Rule of Thirds* because the buildings are placed at the top one-third of the height of the image which makes the image more compelling. The view synthesis generates very dense novel views around the subject which are very difficult to collect manually. The novel views help to better understand the scene and decide the best viewpoint. We believe that future hardware helps to overcome the timing constraint in generating novel and the viewpoint recommendation can be done in real-time.

CONCLUSIONS AND FUTURE WORK

6.1 Conclusions

In this work, we presented an end to end framework to recommend a viewpoint for capturing visually pleasing photographs. Our framework first collects the basis views around the subject of interest and generate camera poses for each basis view. The basis views are used to generate novel views and MPIs, then render novel views using the MPIs. The aesthetic score for each novel views is calculated, the viewpoint with the greatest score is recommended for capturing the photograph. The user survey suggests that 73% of the users agree with the recommended viewpoint for capturing photographs with high aesthetics.

6.2 Limitations and Future Work

The major steps in viewpoint recommendation are extracting camera poses for basis views and generating and MPIs. The camera pose extraction takes about 4-6 minutes and the MPIs generation takes about 5-7 minutes, in total to recommend a viewpoint by processing the complete pipeline it takes 10-15 minutes on an NVIDIA V100 GPU for 30-40 basis views with 32 multi-plane images. Currently, this timing constraint is the major limiting factor for deploying this framework in applications like autonomous photography. However, this could be overcome in future hardware or alternative networks. These observations suggest future work in developing an end to end neural network that can take basis views as inputs and recommends the viewpoints in real-time.

REFERENCES

- [1] Zachary Byers, Michael Dixon, Kevin Goodier, Cindy M Grimm, and William D Smart. An autonomous robot photographer. In *Proceedings 2003 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS 2003)(Cat. No. 03CH37453)*, volume 3, pages 2636–2641. IEEE, 2003.
- [2] Gee-Sern Jison Hsu and Jun-Wei Huang. A photographer robot with multiview face detector. In *2016 IEEE International Conference on Industrial Technology (ICIT)*, pages 2152–2156. IEEE, 2016.
- [3] Kai Lan and Kosuke Sekiyama. Autonomous robot photographer with kl divergence optimization of image composition and human facial direction. *Robotics and Autonomous Systems*, 111:132–144, 2019.
- [4] Kai Lan and Kosuke Sekiyama. Autonomous viewpoint selection of robots based on aesthetic composition evaluation of a photo. In *2015 IEEE Symposium Series on Computational Intelligence*, pages 295–300. IEEE, 2015.
- [5] Ligang Liu, Renjie Chen, Lior Wolf, and Daniel Cohen-Or. Optimizing photo composition. In *Computer Graphics Forum*, volume 29, pages 469–478. Wiley Online Library, 2010.
- [6] Bryan Peterson. *Learning to see creatively: Design, color, and composition in photography*. Amphoto Books, 2015.
- [7] Manfredas Zabarauskas and Stephen Cameron. Luke: An autonomous robot photographer. In *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1809–1815. IEEE, 2014.
- [8] Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016.
- [9] Steven Seitz. Image-based transformation of viewpoint and scene appearance. Technical report, University of Wisconsin-Madison Department of Computer Sciences, 1997.
- [10] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.
- [11] Ben Mildenhall, Pratul P. Srinivasan, Rodrigo Ortiz-Cayon, Nima Khademi Kalantari, Ravi Ramamoorthi, Ren Ng, and Abhishek Kar. Local light field fusion: Practical view synthesis with prescriptive sampling guidelines. *ACM Transactions on Graphics (TOG)*, 2019.
- [12] Peter Hedman, Suhib Alsisan, Richard Szeliski, and Johannes Kopf. Casual 3D Photography. 36(6):234:1–234:15, 2017.

- [13] PE DEBEC. Modeling and rendering architecture from photographs: A hybrid geometry-and image-based approach. In *Proc. SIGGRAPH'96*, 1996.
- [14] John Flynn, Ivan Neulander, James Philbin, and Noah Snavely. Deepstereo: Learning to predict new views from the world's imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5515–5524, 2016.
- [15] Pratul P Srinivasan, Tongzhou Wang, Ashwin Sreelal, Ravi Ramamoorthi, and Ren Ng. Learning to synthesize a 4d rgb-d light field from a single image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2243–2251, 2017.
- [16] Naila Murray, Luca Marchesotti, and Florent Perronnin. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE, 2012.
- [17] Johannes Lutz Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision (ECCV)*, 2016.
- [18] Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.
- [19] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.
- [20] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.
- [21] *Spearman Rank Correlation Coefficient*, pages 502–505. Springer New York, New York, NY, 2008.