Novel-View Synthesis of Human Tourist Photos

Jonathan Freer[1] Kwang Moo Yi[2] Wei Jiang[2] Jongwon Choi[3] Hyung Jin Chang[1]
[1]School of Computer Science, University of Birmingham
[2]University of British Columbia
[3]Department of Advanced Imaging, Chung-Ang University
jxf782@student.bham.ac.uk, {kmyi, jw221}@cs.ubc.ca, choijw@cau.ac.kr, h.j.chang@bham.ac.uk

Figure 1: Provided with a single RGB photo (left) of a known scene, we reconstruct the photo in 3D space in the form of a human mesh and point clouds in a single global coordinate. With this intermediate representation, we then use a deep neural network to re-render the scene from a novel view.

Abstract

We present a novel framework for performing novel-view synthesis on human tourist photos. Given a tourist photo from a known scene, we reconstruct the photo in 3D space through modeling the human and the background independently. We generate a deep buffer from a novel viewpoint of the reconstruction and utilize a deep network to translate the buffer into a photo-realistic rendering of the novel view. We additionally present a method to relight the renderings, allowing for relighting of both human and background to match either the provided input image or any other. The key contributions of our paper are: 1) a framework for performing novel view synthesis on human tourist photos, 2) an appearance transfer method for relighting of humans to match synthesized backgrounds, and 3) a method for estimating lighting properties from a single human photo. We demonstrate the proposed framework on photos from two different scenes of various tourists.

1. Introduction

Imagine you have a photo of yourself where everything is perfect except the camera angle. Perhaps the photographer cut off the top of the attraction you were standing in front of, or you wished the photo had a wider field-of-view. The best course of action to correct this would be to re-take the photo immediately after looking at the outcomes, but this is not always possible, especially for touristic locations, which are often the photos that you care about a lot. Moreover, in situations where there is no "second try", for example when your baby walks for the first time, a re-take is not an option.

With the advent of machine learning, recent works have allowed novel-view re-rendering to be a possibility to resolve this difficulty. Recent works, such as Neural Rendering in the Wild [15] allow the scene to be re-rendered from different cameras. However, they fall short of human rendering leaving room for research in achieving such edits to your everyday photo. Multiplane Imagery (MPI) [11] is another method that could be used for this purpose—they split the view into multiple parallel layers of depth to approximate the parallax. While decent for small viewpoint changes, these methods are not suitable for large camera motion, such as the one shown in Figure. 1.

To overcome the limitations of existing methods, we propose a novel framework for the novel-view synthesis of human tourist photos. Specifically, we propose to model the human and the background separately, where the human is modeled via meshes, and the background via 3D point

clouds. We then combine the two representations together in a novel view to create a deep buffer. This deep buffer is then translated into a realistic rendering of the novel view through a deep network. In order for the rendering to be similar to the original input image, we also utilize the deep features of the original image in the neural rendering process.

In addition to the novel view, we show that our framework can be used to render images of people in different backgrounds, and with different illumination settings. For these renderings to look realistic, it is critical that the lighting is well taken care of. Hence, we further propose to utilize spherical harmonics, as in [12], and introduce how it can be utilized in our neural rendering framework.

To summarize, our contribution is threefold:

- we propose a framework for performing novel view synthesis on human tourist photos;
- we create a method for relighting humans, to match the lighting of synthesized backgrounds;
- we expand on existing lighting estimation work, to show that a simple spherical harmonics estimator is sufficient to provide enhanced lighting coherence.

2. Related Work

Scene Reconstruction / Novel View Synthesis: The ability to synthesize new views of a scene from limited input data is a large area of research in computer vision. Structure from Motion [20] and Multi-View Stereo [21] can be used to reconstruct a sparse point cloud of a scene. The work of Photo Tourism [23] showed how these techniques could be performed on unstructured collections of photographs. Image-based rendering [3][1] methods are able to synthesize new views through the warping of images into 3D geometry, however rely on dense captures of the scene. Multiplane Images [11] allow view synthesis from a single view by applying depth estimation to slice the input image into multiple planes which are projected away from the camera. However, the movement of the camera is limited due to the need to inpaint [9] areas and the 2-dimensional nature of the planes revealing themselves.

Neural Rendering: Over the past few years, there has been extensive research into using Neural Rendering for novel view synthesis, with several recent approaches employing volume rendering methods. Neural Radiance Fields (NeRF) [16] utilizes a multi-layer perceptron (MLP) to model radiance fields, allowing for photo-realistic renderings of novel viewpoints and consistent reconstruction even with large camera motions. NeRF in the Wild [13] applies this method to complex scenes using unstructured collections of in-the-wild photographs. Other approaches rely on image translation networks to re-render traditionally rendered reconstructed scenes, such as dense point clouds. Neural Rerendering in the Wild [15] takes in as input a deep buffer containing a rendered point cloud, a semantic mask, and a latent appearance vector and produces a realistic rerendering of the provided point cloud, with the ability to vary appearance such as season and time of day.

Human Synthesis: Many techniques have been proposed to synthesize imagery of humans from alternate views. Person image synthesis works [26][14] use Generative Adversarial Networks (GANs) [5] to repose humans, by taking in an input image of a human and a target pose. Unfortunately, these methods fail to preserve identity when the input data is too dissimilar to the training data. Recent works [18] have employed Neural Radiance Fields [16] to reconstruct a human scene from numerous input images. Human Digitization works, such as PIFu [19], present methods for reconstructing textured meshes from single image inputs, allowing for the rendering of the model from alternate views.

Lighting Estimation: Light probes are a common form of representing real-world lighting in computer graphics. Traditionally, a set of cameras are used at a location to capture the lighting information, forming an active light probe. Deep Outdoor Illumination Estimation [8] shows how a CNN can estimate high-dynamic-range illumination from a single image. Learning Lightprobes [12] presents a method for estimating light probes to allow for lighting in mixed reality applications. Learning Lightprobes employs a series of CNN's to learn the lighting of a single object from independent angles, representing the lighting as spherical harmonics [6]. Relighting Humans [10] presents methods for accurately relighting humans provided, through the use of a CNN to infer light transport and albedo maps, from a single image.

3. Methodology

Due to the tendency of existing neural rendering methods to segment transient objects such as humans in the training data in order to discourage their rendering and the artifacts that come with them, these methods tend to perform poorly when required to render humans. Our method addresses this by separating the task of human and scene rendering into two sections, each handled by an independent set of networks. This allows for photo-realistic rendering of both human and scene simultaneously, with minimal artifacts.

3.1. Framework

Figure 2 shows the overall framework of the method. Our framework comprises three main components: a) Human Digitization and Localization, b) a Traditional Renderer and c) a Neural Renderer. For Human Digitization and

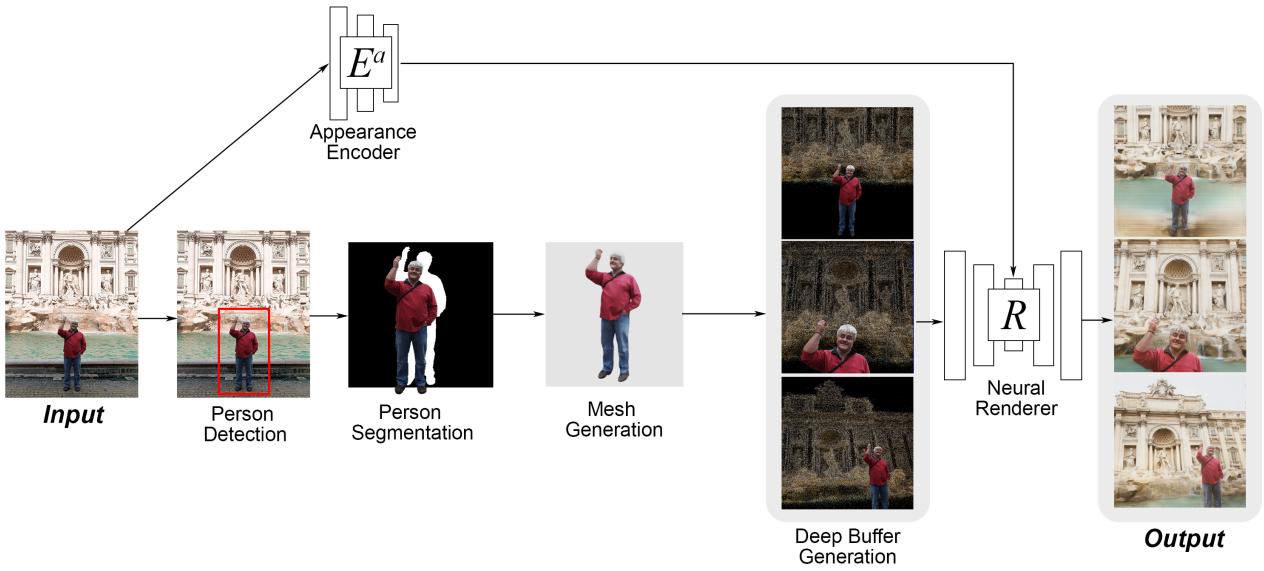

Figure 2: Overview of our method. Given a photo, person detection, segmentation, depth estimation, and mesh generation is performed using off-the-shelf algorithms. The generated mesh is textured and combined with a scene point cloud. The point cloud mesh combination can then be rendered from novel views into an aligned deep buffer (consisting of depth, color and semantic labeling). An appearance vector is generated from the Appearance Encoder $\mathbf{E^a}$ using the input image as input. The Neural Renderer **R** takes a combined input of the appearance vector and the deep buffer, producing the photo-realistic output.

localization, we use a combination of pre-trained off-the-shelf models for person segmentation, mesh reconstruction, and depth estimation. For Neural Rendering, we expand on the framework presented in Neural Rerendering in the Wild [15] to allow for human pass-through rendering and human appearance transfer. We use traditional rendering techniques to combine the previously modeled human and background and generate a deferred-shading deep buffer input for the Neural Renderer.

3.2. Human Digitization and Localization

In this component, we aimed to solve the problem of reconstructing the mesh of a person and estimating their relative position to the camera so that they could later be placed back into the scene with correct scales. To locate all humans in the input image, the Mask R-CNN [7] network was used. We found that the semantic segmentation from this network was overly smoothed, resulting in poorly shaped reconstructed human meshes. To improve the segmentation, we use Mask R-CNN solely as person detection and use the output of this network to feed the DeepLab [2] semantic image segmentation network trained on ADE20K [25]. This was found to produce crisp segmentation, without noise and over smoothing, compared to using each of the networks independently. To reconstruct the human mesh, we use the PIFu [19] Network with pre-trained weights.

To estimate the distance between the human and the camera, we employ the MonoDepth2 [4] monocular depth estimation network. Due to lack of linearity in the estimated distances and the need to align the depth estimation coordinate system to the pre-existing point cloud scene coordinate system, we learned a scaling by comparing the predicted depth of buildings in the scene to a true depth generated by COLMAP [22][21][20]. We apply Image Registration [17] through COLMAP to estimate the intrinsic and extrinsic of the cameras. Combining the extrinsic camera properties, the predicted depth, and the human pixel offset: the human mesh can be projected into the scene away from the camera and be correctly placed and scaled in the 3D scene coordinate space.

3.3. Deep Buffer Generation

We utilize OpenGL [24] to generate an aligned deferred-shading deep buffer, containing pixel albedo, depth map, and binary segmentation map. During this stage we project the segmented human pixel information back onto the reconstructed mesh, to improve the texture quality of the mesh. Additionally in this stage, a number of pre-processing steps can be performed, such as normalizing the human pixels, or applying any directional spherical harmonic lighting. For directional lighting estimation and application, provided with an appearance reference image containing a human, we estimate the lighting coefficients for that human through the use of the Lighting Estimation network. This lighting is then applied onto the mesh in the rendering stage. Finally we render the point cloud together

with the generated human mesh, to produce a single pixel albedo layer. We produce the depth map by performing a separate rendering pass on just the point cloud, altering the color values of each point to reflect the distance from the camera.

3.4. Neural Renderer

We utilize a deep network Neural Renderer to transfer the deep buffer into realistic output images. Our Neural Renderer is built on top of that of Neural Rerendering in the Wilds (NRiW). Similar to that of NRiW we generate an aligned dataset of reference, color, depth, and segmentation images. The segmentation map which was introduced in NRiW, aimed to reduce neural artifacts from transient objects, such as humans and other non-permanent objects, being learned into the scene. Due to the inability to synthesize a segmentation map from novel views, we train our network on a binary human and non-human segmentation map. This allows our network to distinguish the human pixels, which need to be passed through the network, from the point cloud which needs to be transferred into a photo-realistic representation. We found that the use of a binary segmentation map still managed to eliminate much of the neural artifacts from transient objects.

To train the network to pass-through the human pixels from the input data, during training we transfer human pixels from the reference image directly into the corresponding rendered image. During this process, we normalize the transferred pixels by subtracting the average color intensities of each human. Through this we aim to eliminate the lighting characteristics on each human, allowing the network to learn to relight the humans to match the reference image based on the corresponding appearance vector.

We generate a deferred-shading deep buffer, containing pixel albedo, the depth map, and binary segmentation map. We render the point cloud together with the generated human mesh, to produce the pixel albedo layer. Additionally, in the rendering, we normalize and randomly drop the human pixels. We randomly drop transferred pixels, to encourage the network to learn missing data, in order to be able to improve the quality of rendering when low-resolution human data is provided from the Human Digitization component.

3.5. Lighting Estimation

The work of Learned Lighprobes [12] presents a method for estimation of spherical harmonic lighting coefficients for a fixed model from a fixed angle. We expand upon this work to allow for spherical harmonic lighting estimation on any human independent of angle, from a single image.

To achieve this we designed a dataset of images of humans with known lighting characteristics. We generated a dataset consisting of 200 photos of individual humans, designed to cover a diverse spectrum of clothes, race, and gender. Each of the 200 images and the corresponding masks, generated using DeepLab [2] Semantic Image Segmentation, were used as input data into the PIFu [19] network, producing corresponding meshes. We relight each of the meshes using spherical harmonic lighting methods [6], to produce a dataset of humans with known lighting properties. We found fixing 15 coefficients in a binary state (low and high) able to represent a high level of illumination resolution in a limited amount of data. For each model, one-tenth of the 2^{15} different coefficient combinations were randomly sampled and rendered, resulting in a dataset **T** containing 655,000 images.

We trained our CNN using **T** as training data and the known lighting as reference for the loss function. Our CNN is implemented on top of that of Learned Lightprobes [12], taking in an image of a human and outputting 15 estimated coefficients. We found the loss function of mean squared error across the coefficients to perform poorly, due to differences in coefficients not correlating with differences in illumination. To solve this we utilize a sample-based loss function, where both the true illumination and the estimated illumination are projected onto spheres and the MSE across the two spheres becomes the loss. We found sampling at 500 points on each sphere to provide an accurate representation of the differences in illumination. We pre-computed a matrix representing each of the points on the sphere and their spherical harmonic amplitudes associated with the angle of the normal, allowing for the loss to be calculated using only a single matrix multiplication for each sphere. During train time, the 500 (points) x 15 (coefficients) matrix of constants would be loaded into memory and used in the loss function of each operation.

4. Experiments

4.1. Dataset

We trained our method on two datasets reconstructed with COLMAP [22] from public images, Trevi Fountain (3006 images) and the Brandenburg Gate (1363 images), chosen for covering a wide range of appearances, the composition of the scenes, and occluding transient objects. A separate model is trained for each dataset. We create an aligned dataset by rendering the reconstruction twice for each reference image. Similar to Neural Rerendering in the Wild, we render an output corresponding to a 512x512 center crop of the reference image. However, we found that this by itself performed poorly when wide-angled uncropped images were evaluated on this method. To alleviate this we render an additional wide-angle output for each reference image and downsample the reference image instead of cropping. This allows us to cover a much greater range of focal lengths in training from the same size dataset. We

Figure 3: Examples of image reconstruction and novel view rendering using our framework. Our method can render highly realistic images from a wide viewing angles and maintain the lighting condition.

Table 1: Average error on the validation set using VGG/perceptual loss (lower is better), L_1 loss (lower is better), and PSNR (higher is better).

VGG	L_1	PSNR
0.484	41.54	12.38

form the validation set by randomly selecting 100 images per dataset.

For evaluation, as our framework requires full human bodies for the Human Reconstruction and Localization, we create a novel dataset of images from the original datasets which meet this requirement.

4.2. Reconstruction Metrics

To report image reconstruction errors in the validation set we used perceptual loss, L1 loss, and PSNR. We compare the ground truth input image, to the rendered view from the same camera angle. Due to the novel nature of our work and the input requirements of our framework, our work is not directly comparable with other works in the area, so we have reported quantitative results independently. The metrics can be seen in Table 1.

4.3. Scene Reconstruction

Figure 3 demonstrates the reconstruction of images using our framework. They show a realistic rendering of two separate scenes from three unique viewpoints each. In all output images, neural artifacts can be found in areas where the point cloud input is sparse. In the figure Brandenberg Gate Scene (second row) dark neural artifacts can be seen in the sky, a result of non-human transient objects being learned into the scene. This can be corrected by using a three-part segmentation map in the training stage, segmenting humans, non-human transient objects such as temporary barriers and signage, and everything else. During reconstruction a binary human/ non-human segmentation map can still be used, resulting in transient artifacts not being rendered. Figure 4 shows how a human can be transferred from one scene to another, realistically recreating images as if they were present at the location.

4.4. Appearance Transfer

Figure 5 shows how appearance can be transferred from a separate image to the output. It demonstrates realistic relighting of the human in the foreground to match the appearance of the rendered background. The figure shows that while hue and brightness are transferred, directional lighting such as specular is not transferred. This is a result of uniformly normalizing the human pixels in the input data as

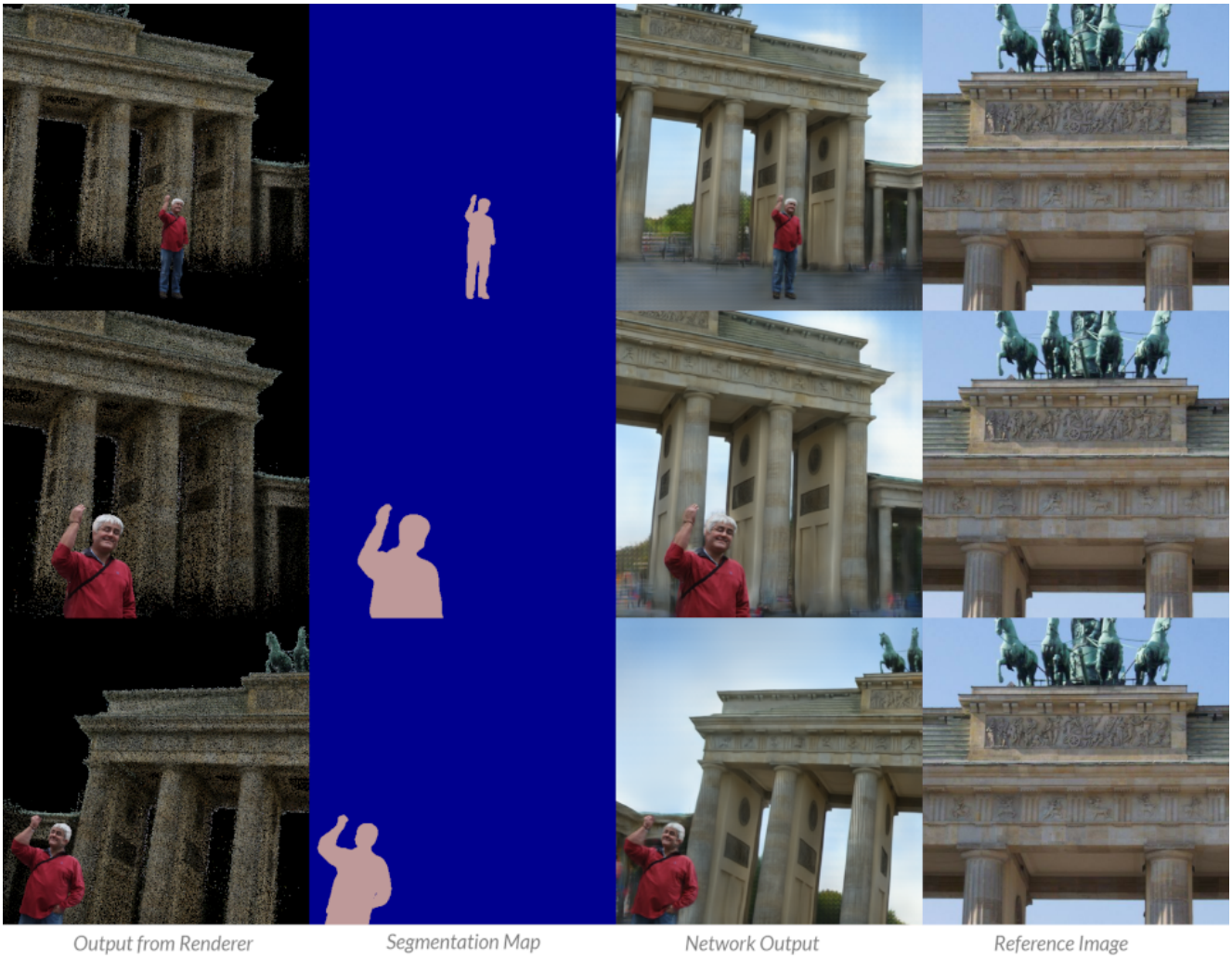

Figure 4: Example of human transferred from one scene to another. Here a human from the Trevi Fountain Dataset is transferred into the Brandenburg Gate scene. From left to right: Output from Traditional Renderer, Binary Segmentation Mask, Framework Output, Appearance Reference Image.

opposed to reconstructing the albedo.

4.5. Lighting Estimation

We evaluate our lighting estimation method by comparing the difference in true illumination intensity to estimated intensity, through the sampling of intensity when applying the spherical harmonic lighting to a sphere. Figure 6 shows the results for lighting estimation. The network is able to estimate the lower level harmonics of the illumination well, however struggles to match the high-level details. While the network is able to estimate lighting behind the human, it performs far better on lighting in front of the human. The smoothed nature of the estimated lighting can be attributed to pre-existing lighting on humans before the known lighting was applied. To alleviate this the albedo texture for the human mesh should be synthesized and during lighting application.

4.6. Limitations

We note the following limitations of our framework: (1) The requirement for a full human body to be present in the input for mesh reconstruction to work. (2) Inaccuracies in the segmentation and mesh generation networks, may result in a person which looks unrealistic. (3) Inaccuracies in-depth estimation will result in the human mesh being placed and scaled incorrectly in the scene. (4) A number of different neural artifacts may be present in the final render. Phantom objects, such as blurs, may appear where transient objects have been learned into the scene. Areas that were not frequent in the training data or where limited point cloud

Figure 5: Example of appearance transfer from real images. The top row represents the source appearance real image used as the input to the appearance encoder. The bottom two rows represent the out from the rendering network with appearance transferred onto two unique views.

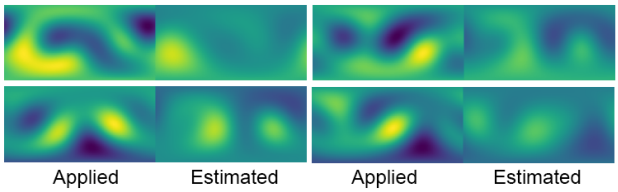

Figure 6: Lighting Estimation: Example of applied lighting intensity (left) against predicted lighting intensity (right). Note how whilst the overall intensity is not matched, the model is still able to estimate the pattern of the lighting.

information is provided, such as the ground, fail to render accurately. This limits the possible views to only those that focus on material featured in the training data. (5) The requirement for a photo dataset of the location the photo was taken from, required to generate a point cloud and train the neural renderer.

5. Conclusion

This paper has presented a method to recreate novel views of single-person tourist photos. Our method allows for a wide range of motion away from the original viewpoint, while still rendering realistic imagery. We train a deep network to enable appearance transfer from reference images to both human foreground and static backgrounds, allowing for a time of day or season adjustments to novel renders. Furthermore, we show that our framework can be used to render images of people in different backgrounds, and with different illumination settings. Our experimental results show that the proposed method is able to produce photo-realistic renders of novel views.

We believe our method presents a wide range of possibilities, from simply producing novel views to recreating whole video sequences for CG work. We hope to expand this work to allow interactive photo-realistic scenes, through rigging on the human mesh, to allow custom animations or user control.

Acknowledgement: This work was partially supported by systems provided by Compute Canada, by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant, and by the Institute of Information and communications Technology Planning and evaluation (IITP) grants funded by the Korea government (MSIT) (No.2021-0-00034, No.2021-0-00537, No.2021-0-01341, and Artificial Intelligence Graduate School Program (Chung-Ang University)).

References

[1] Chris Buehler, Michael Bosse, Leonard McMillan, Steven Gortler, and Michael Cohen. Unstructured lumigraph rendering. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, SIGGRAPH '01, page 425–432, New York, NY, USA, 2001. Association for Computing Machinery.

[2] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *CoRR*, abs/1606.00915, 2016.

[3] Yasutaka Furukawa and Jean Ponce. Accurate, dense, and robust multiview stereopsis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 32(8):1362–1376, 2010.

[4] Clément Godard, Oisin Mac Aodha, and Gabriel J. Brostow. Digging into self-supervised monocular depth estimation. *CoRR*, abs/1806.01260, 2018.

[5] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[6] R. Green. Spherical harmonic lighting: The gritty details. 2003.

[7] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. Mask R-CNN. *CoRR*, abs/1703.06870, 2017.

[8] Yannick Hold-Geoffroy, Kalyan Sunkavalli, Sunil Hadap, Emiliano Gambaretto, and Jean-François Lalonde. Deep outdoor illumination estimation. *CoRR*, abs/1611.06403, 2016.

[9] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Trans. Graph.*, 36(4), July 2017.

[10] Yoshihiro Kanamori and Yuki Endo. Relighting humans: Occlusion-aware inverse rendering for full-body human images. *CoRR*, abs/1908.02714, 2019.

[11] Diogo C. Luvizon, Gustavo Sutter P. Carvalho, Andreza A. dos Santos, Jhonatas S. Conceição, Jose L. Flores-Campana, Luis G. L. Decker, Marcos R. Souza, Hélio Pedrini, Antonio Joia, and Otávio A. B. Penatti. Adaptive multiplane image generation from a single internet picture. *CoRR*, abs/2011.13317, 2020.

[12] David Mandl, Kwang Moo Yi, Peter Mohr, Peter M. Roth, Pascal Fua, Vincent Lepetit, Dieter Schmalstieg, and Denis Kalkofen. Learning lightprobes for mixed reality illumination. In *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 82–89, 2017.

[13] Ricardo Martin-Brualla, Noha Radwan, Mehdi S. M. Sajjadi, Jonathan T. Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections, 2021.

[14] Yifang Men, Yiming Mao, Yuning Jiang, Wei-Ying Ma, and Zhouhui Lian. Controllable person image synthesis with attribute-decomposed gan. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[15] Moustafa Meshry, Dan B. Goldman, Sameh Khamis, Hugues Hoppe, Rohit Pandey, Noah Snavely, and Ricardo Martin-Brualla. Neural rerendering in the wild. *CoRR*, abs/1904.04290, 2019.

[16] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision – ECCV 2020*, pages 405–421, Cham, 2020. Springer International Publishing.

[17] Sayan Nag. Image registration techniques: A survey. *CoRR*, abs/1712.07540, 2017.

[18] Keunhong Park, Utkarsh Sinha, Jonathan T. Barron, Sofien Bouaziz, Dan B Goldman, Steven M. Seitz, and Ricardo Martin-Brualla. Nerfies: Deformable neural radiance fields. *ICCV*, 2021.

[19] Shunsuke Saito, , Zeng Huang, Ryota Natsume, Shigeo Morishima, Angjoo Kanazawa, and Hao Li. Pifu: Pixel-aligned implicit function for high-resolution clothed human digitization. *arXiv preprint arXiv:1905.05172*, 2019.

[20] Johannes L. Schonberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[21] Johannes Schönberger, Enliang Zheng, Marc Pollefeys, and Jan-Michael Frahm. Pixelwise view selection for unstructured multi-view stereo. volume 9907, 10 2016.

[22] Johannes L. Schönberger. Colmap, 2016. `http://colmap.github.io/`.

[23] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3):835–846, July 2006.

[24] Mason Woo, Jackie Neider, Tom Davis, and Dave Shreiner. *OpenGL programming guide: the official guide to learning OpenGL, version 1.2*. Addison-Wesley Longman Publishing Co., Inc., 1999.

[25] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5122–5130, 2017.

[26] Zhen Zhu, Tengteng Huang, Baoguang Shi, Miao Yu, Bofei Wang, and Xiang Bai. Progressive pose attention transfer for person image generation. *CoRR*, abs/1904.03349, 2019.