



Illumination-aware group portrait compositor

Masaru Ohkawara¹ · Issei Fujishiro¹

Accepted: 20 April 2022 / Published online: 20 May 2022

© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2022, corrected publication 2022

Abstract

We present a novel compositing framework for full-length human figures that maintains their surface details and appends the localized nature of light and shadow, thereby synthesizing composite results with high visual coherence. The framework is extended from the compositing pipeline proposed in our previous study so that it deploys five stages for photometric information estimation, as well as for 3D reconstruction, global illumination simulation, lighting transfer, and compositing. Based on the interpretation that a sense of coexistence can be achieved through visual coherence, we demonstrate that the proposed framework functions properly as a group portrait compositor. The composite results that the proposed framework composed the images separately rendered 3D human models compared favorably with the results which rendered multiple avatars together. Based on this empirical evaluation, the proposed framework is expected as a new means of fostering a sense of coexistence in remote societies and of efficiently generating highly photorealistic cyberworlds.

Keywords Image processing · Digital image compositing · Visual coherence · Perception

1 Introduction

Image compositing, which combines multiple visual elements to generate a single image, is known as one of the fundamental techniques of visual effects. In particular, compositing human figures onto a specific scene to create the illusion of being there has a wide range of applications, including virtual scenes in movie production [6], video conferencing [34], and portrait shooting/editing on smartphones [39].

Compositing different visual elements in a harmonized manner is considered an important task in image compositing. Early studies in this field proposed graphcut textures [17] and Poisson image editing [21] as blending methods focusing on image gradients. In seamless image cloning based on solving Poisson equation, an extended method [11] has also been proposed. The quality of these methods is unfortunately limited because they rely only on the image space information, not 3D illumination information. Once the method of image-based lighting [9] was established, relighting methods that use information from off-screen space, i.e., areas outside the image, were proposed. With the development of deep learning, the quality of compositing has continued to be enhanced

with improvements in off-screen space estimation. However, most compositing studies focus on relighting. For example, the work proposed by Xu et al. [38] can shade the object that will be inserted in a scene based on lighting estimation, while the shadow is cast by just one dominant directional light. That is, the effect of shadows caused by localized occlusions is not considered. This limits the scope of applications only to cases where the effects of shadows can be ignored, such as bust shots or scenes with no elements other than people. On the other hand, our previous study [19] incorporated 3D reconstruction and global illumination simulation, which are not handled in conventional relighting methods, to enable high-quality compositing in situations where the effects of shadows caused by localized occlusions cannot be ignored.

In this paper, we extend the compositing pipeline proposed in our previous study [19] to demonstrate that the proposed framework functions properly for its new example application: group portrait compositing. Group portraits are obtained by capturing several people that share the same time and space in a common lighting environment. In group portraits, all the elements should have the same color tone, and the existence of the elements should be optically constrained by the shadows cast onto the elements and surrounding structures. These features give an irreplaceable sense of coexistence to group portraits.

✉ Masaru Ohkawara
masaru.ohkawara@fj.ics.keio.ac.jp

¹ Keio University, Yokohama, Japan

The global response to the COVID-19 pandemic has led to many remote collaborations. Owing to the shift to the era of “remote everything,” many people feel less connected to others and more fatigued with few face-to-face collaborations [15]. For example, Microsoft Teams aims to address this issue with an additional function called Together Mode [33], but this function is only a simple collection of snapshots, and there remains room for improvement in terms of visual coherence. The proposed framework is intended to reproduce a sense of coexistence, an intrinsic value of group portraits, by relighting people captured in different lighting environments under a new lighting environment and by casting shadows that account for the occlusions that occur between each element, as illustrated in Fig. 1. The ability to produce high-quality composite results of multiple people gathering is expected to establish a new means of fostering a sense of coexistence in remote societies and of efficiently generating highly photorealistic cyberworlds.

The main contributions of this paper are twofold:

- (1) We establish a novel lighting representation within a neural network updated from our previous study [19], which translates humans’ surface normals to their high dynamic range diffuse reflection components, to utilize them for state-of-the-art lighting transfer. It considers both the relighting of full-length humans and the reprojecting of their shadows.
- (2) We demonstrate that the proposed framework functions properly for its new example application: group portrait compositing, and we demonstrate the ability of the proposed framework to create group portraits in remote societies.

The remainder of this paper is organized as follows. Section 2 introduces related works. Section 3 presents our revised compositing framework. Section 4 demonstrates that the proposed framework functions properly as a group portrait compositor. Section 5 discusses the limitation of the proposed framework, and Sect. 6 concludes the paper with some directions for future research.

2 Related works

The proposed method builds on knowledge from several domains. First, Sect. 2.1 describes conventional image-based relighting methods, and Sect. 2.2 reviews the related works for shadow editing, which is particularly important in image compositing. Further, Sect. 2.3 reviews some aspects of image-to-image translation. Finally, Sect. 2.4 highlights our approach in terms of differences from related works.

2.1 Image-based relighting

The origin of global illumination is believed to be environment mapping [5], which is a kind of texture mapping that refers to texels from the surface of a cube or sphere surrounding a target object. Lighting methods using environment mapping are referred to as image-based lighting [9], from which image-based relighting was derived. Debevec et al. used one-light-at-a-time (OLAT) images to obtain reflectance fields, which enable the relighting of the human face in any desired environment [10]. Recently, many methods based on deep learning have been proposed. Sun et al. realized relighting by estimating an environment map from a given image and retargeting it to the desired environment map [32]. Kanamori et al. estimated the albedo, light transport map, and ambient light to achieve the relighting of full-length humans by considering their self-shadows [16]. However, they focus only on humans as the composite target and do not consider the effects of humans on other elements, including cast shadows.

2.2 Shadow editing

Shadows play an important role as visual cues in human perception and thus are vital for convincing image compositing. Chuang et al. enabled shadow reprojection for complex shapes with the shadow compositing equation [8]. However, the work does not support relighting and makes it hard to edit flexibly the direction and color of shadows. There is also a known method to reproduce optically correct shadows while image compositing [7], but it has some limitations, one of which is that this method is only valid for scenes with one dominant, high directive light source, that is, it does not support global illumination effects. Wang et al. developed a generative adversarial network (GAN)-based framework to estimate and cast the shadows of a composite target from a given background static video sequence [37]. However, due to the characteristics of GAN, it is difficult to edit lighting conditions flexibly. Philip et al. reconstructed proxy shapes from multi-view images to enable shadow removal and reprojection while relighting [22], and our framework was strongly inspired by this flow of relighting via proxy shapes.

2.3 Image-to-image translation

Color transfer [26] is a well-known method for matching the color statistics of source and target images. It enables the source image to be transferred linearly to another image with a color tone similar to that of the target image. This method was later extended to support nonlinear transfers [23], making it possible to perform more general transfers. There exists a method to transfer lighting for human face relighting with a single source image and a single reference image [18]. A

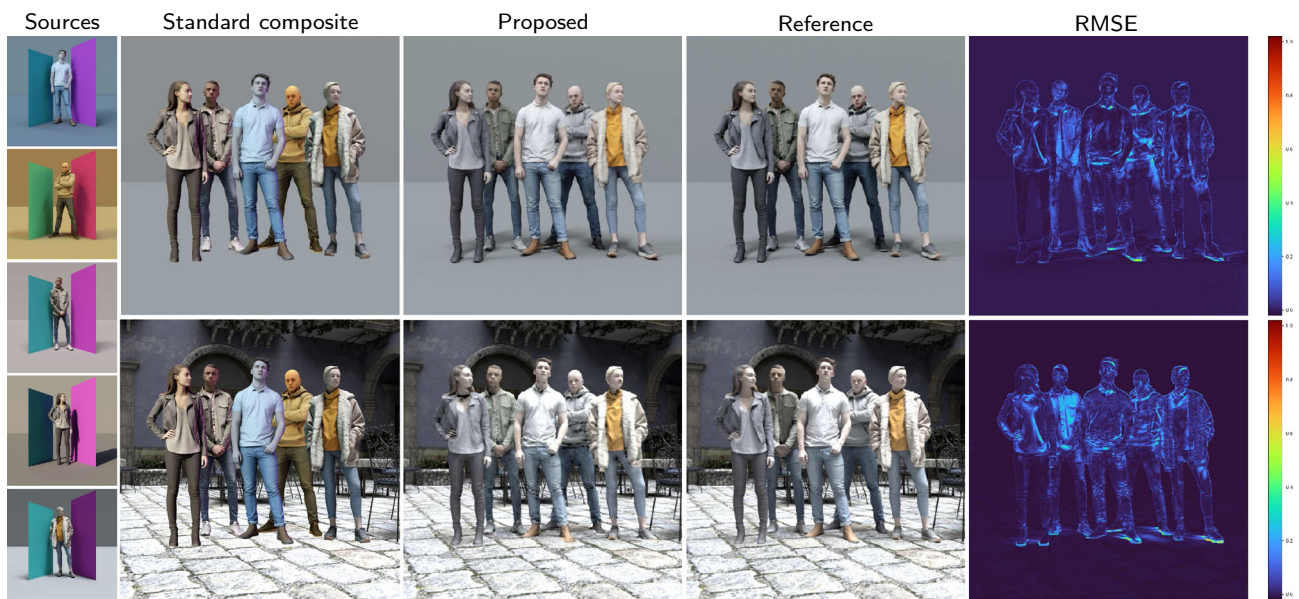


Fig. 1 Composite group portraits. The group portraits were produced from five individual portraits shown in the column Sources, each of which was captured in a different environment. Standard composite shows the composite results without considering visual coherence. Proposed shows the composite results using the proposed framework. Reference shows the rendered images of the 3D reference models. RMSE visualizes the root mean square error for each pixel between Proposed and Reference with a Turbo colormap [2], which considers visual con-

tinuity. The composite images without considering visual coherence have neither a uniform color tone among the elements nor cast shadows, while the composite images produced by the proposed framework compare favorably with the rendered one. Further, the upper portrait is configured with a simple background scene and the visual effects between the people and the scene can be confirmed clearly, while the lower one is configured with a photorealistic background scene and it evokes actual usage scenarios

recent method achieved a more geometry-aware transfer by adding geometric properties [31]. In deep learning, many image-to-image translation networks consisting of encoders and decoders have been proposed, and each has achieved remarkable results. Among them, pix2pix [14] is a GAN-based method that incorporates U-net [28] as a generator, and it is highly versatile. In terms of relighting, there is a translation method [35] to harmonize the colors of the foreground and background elements in a composite image, but it does not consider 3D lighting.

2.4 Our approach

We present a novel compositing framework that considers visual coherence. The proposed framework reproduces shadows caused by localized occlusions. Note that most existing image-based methods are formulated on environment lighting, which can consider only illumination from a textured surface on an infinite distant sphere/cube, so they cannot handle localized occlusions. Because localized occlusions are likely to occur in the case of group portrait compositing, a method to bridge the 3D and 2D information is needed. In the proposed framework, we utilize a single-view 3D reconstruction method to reconstruct a low-resolution proxy shape and to simulate global illumination to obtain more precise illumi-

nation properties. The low-resolution shape is then rendered into a 2D image, and by mapping its illumination properties to the source image, the illumination properties are transferred to the composite target. This series of processes enables inter-reflection between the composite target and its surrounding elements, which could not be represented using conventional methods.

3 Framework

Figure 2 schematically illustrates our framework. For brevity and simplicity, a case of individual portrait compositing is described in this section. The photometric information estimation stage estimates the alpha matte, albedo, and surface normals of a full-length human figure from a given RGB image. The shape reconstruction stage reconstructs the entire human model from a single-view image. The global illumination simulation stage uses the reconstructed shape and the given 3D scene to output globally illuminated images for the background and foreground images separately. Here, the two globally illuminated images include photometric information, i.e., alpha matte, albedo, and surface normals of the reconstructed shape. The lighting transfer stage maps the surface normals and diffuse reflection components to the

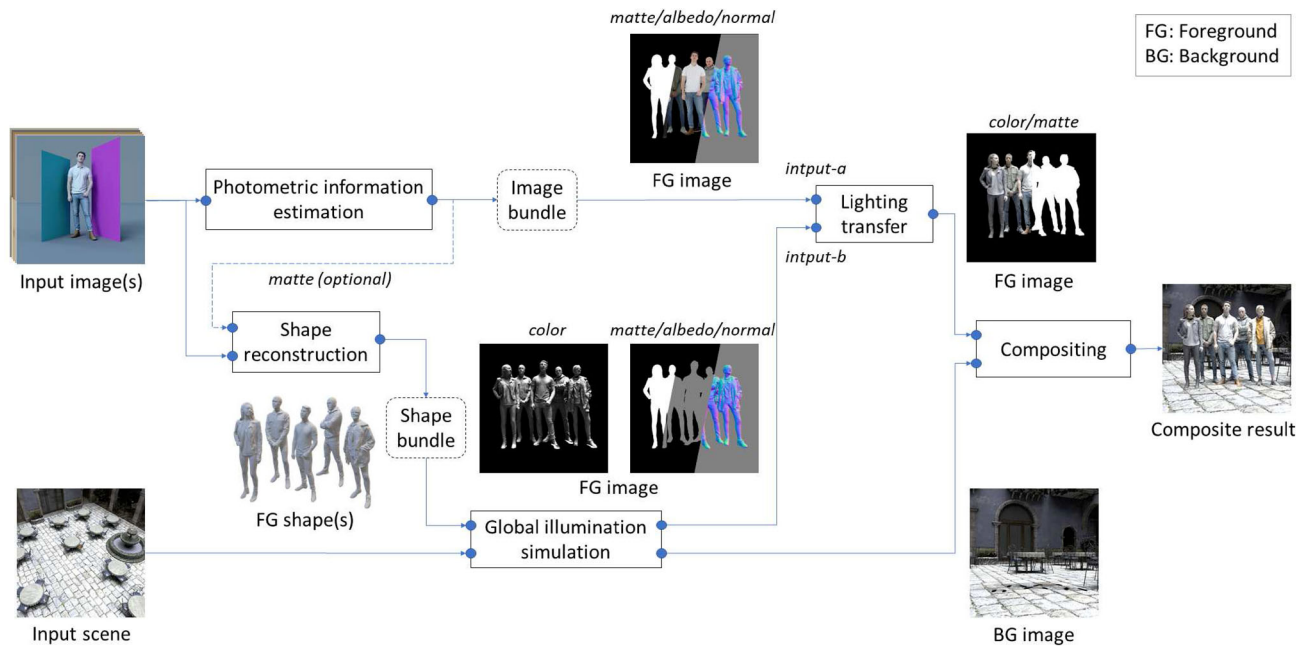


Fig. 2 Proposed framework. The proposed framework starts with the photometric information estimation and 3D reconstruction for a single full-length human figure. The reconstructed shape(s) is placed onto a given 3D scene manually and global illumination is simulated; then, the globally illuminated images for the background and foreground are output separately. The lighting is retargeted to the estimated photomet-

ric information (input-a) based on the obtained photometric information (input-b). Finally, the retargeted foreground image is superimposed onto the background image and converted through tone mapping from high dynamic range to standard dynamic range to produce the composite result

globally illuminated images, and applies this map to the estimated surface normals at the photometric information estimation stage to transfer the diffuse reflection components and achieve relighting. Finally, the compositing stage superimposes the foreground image onto the background one and converts through tone mapping the dynamic range of the composite image from high dynamic range (HDR) to standard dynamic range (SDR).

The 3D human models and the HDR environment maps used in our experiments were acquired from 3D Scan Store¹ and Poly Haven², respectively. A laptop computer with an Intel® Core™ i9-9980HK CPU @ 2.4GHZ (8 cores, 16 threads), 64 GB memory, and a NVIDIA GeForce RTX 2080 with Max-Q Design GPU (8 GB VRAM) was used for the system implementation and all the experiments.

3.1 Photometric information estimation

Because our previous study [19] cannot be applied to in-the-wild portraits—portraits whose 3D information is unavailable—we needed to estimate the foreground alpha

matte, albedo, and surface normals of a full-length human figure from a given RGB image. Then, a stage for estimating various photometric information was added.

Preliminary experiments have revealed that high accuracy is required for each estimation at this stage. Figure 3 compares the composite results, one of which uses the ground truth photometric information and the other uses the estimated photometric information. To estimate the alpha matte, albedo, and surface normals, U²-net [25], I1W [4], and pix2pixHD [36] were employed, respectively. Figure 3 shows that even if the highest-level methods currently available are applied, the photometric information for obtaining a satisfied composite result cannot be estimated. To address this issue effectively, a known deep learning-based method [20] using OLAT images is considered useful. If the pre-trained model of this method was available, high-quality alpha matte, albedo, and surface normals could have been estimated from a single-view image. Note that the pre-trained model is unfortunately unavailable at this time, so the alpha matte, albedo, and surface normals are assumed given in this paper.

3.2 Shape reconstruction

PIFuHD [30] is known as a state-of-the-art method to reconstruct a highly detailed entire human model from a

¹ <https://www.3dscanstore.com/3d-model-bundle/arch-viz-mega-bundle>.

² <https://polyhaven.com/hdris>.

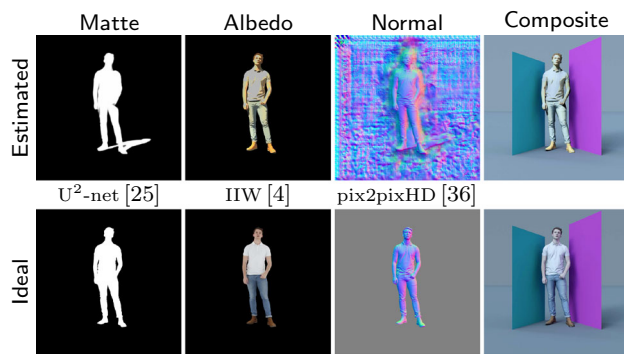


Fig. 3 Low-quality photometric information leads to poor composite results. The composite result that uses the estimated photometric information is compared with the other composite result that uses the ground truth photometric information. This preliminary experiment shows that even if the highest-level methods currently available are applied, the photometric information for obtaining a satisfied composite result cannot be estimated. U²-net [25], IIW [4], and pix2pixHD [36]

single-view image. Its pre-trained model is available and can be adopted to digitize humans from full-length portraits. At this stage, we decided to reconstruct shapes with the $512 \times 512 \times 512$ grid cells, which is the prescribed resolution in the paper [30]. It is also shown there that PIFuHD has extremely high accuracy compared with competing methods, such as Tex2Shape [1], PIFu [29], and DeepHuman [40], but when comparing the reconstruction results with the ground truth, the details are not completely reproduced, contrary to our expectation, as demonstrated in Fig. 4. In addition, the pre-trained models do not include texture restoration. Therefore, direct rendering of the reconstructed shape will not result in complete visual content. Here, the reconstruction target will be clarified in terms of human poses, camera angles, and props. First, the supported human poses are limited to standing poses. Because the pretrained model of PIFuHD is fitted to the clothed human dataset, it mostly consists of standing poses. However, ARCH++ [12] has recently succeeded in reconstructing sitting poses with the same dataset as PIFuHD. This implies that target poses can be diversified by changing the 3D reconstruction method. Second, the camera angle will be desirable to fix the elevation 0 degrees, because the pretrained model of PIFuHD is fitted to the images rendered with the elevation fixed at 0 degrees. The images that deviated from this condition are more likely to fail the reconstruction. Finally, the props are partially supported. In our preliminary experiment, the person who carried their bag was totally reconstructed, but there is the failed case in a paper on PIFuHD [30]. This is due to unavailability of large-scale and diverse avatars. On the other hand, the props surrounding the people can be composed by creating them as part of the background scene.

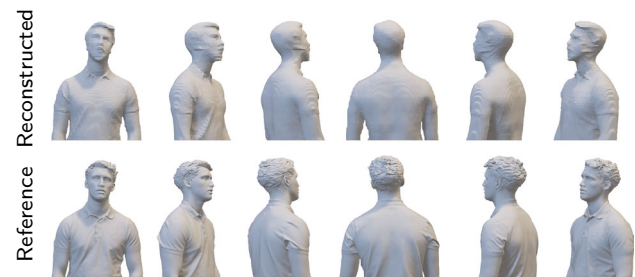


Fig. 4 Comparing the reconstructed shape using PIFuHD with the ground truth 3D model, the reconstructed shape lacks details

3.3 Global illumination simulation

The reconstructed shape is placed onto the given 3D scene manually, and then the global illumination simulation begins. Autodesk Maya (ver.: 2022) and Arnold for Maya (ver.: MtoA 4.2.4) were employed as the digital content creation tool and renderer, respectively. In the global illumination simulation, the alpha matte, albedo, and surface normals were rendered as arbitrary output variables (AOVs), in addition to shading color using Arnold with GPU acceleration. Here, AOVs are secondary images generated by renderers. Any number of AOVs can be produced simultaneously, and each may be used to edit the corresponding lighting component while compositing. Most renderers support AOVs' output, including Arnold [3]. Because the reconstructed shape has neither the material nor the texture, we set its material as a Lambertian model and its albedo as $(H, S, V) = (0.0, 0.0, 0.5)$. In the previous study [19], negative values of surface normals were clamped, but herein, the range of values was converted from $[-1, 1]$ to $[0, 1]$, which successfully improved the prediction accuracy of the correspondence between the surface normals and the diffuse reflection component at the lighting transfer stage.

3.4 Lighting transfer

The structure of the lighting transfer module is illustrated in Fig. 5. The illumination net predicts the diffuse reflection component from the alpha matte and surface normals. The illumination net plays a central role in translating the shading color from a low-detailed shape to high-detailed one. Here, this translation holds the assumption that the low-detailed and high-detailed surface normals are well correlated and the correspondence between the surface normals and shading color is invariant with respect to level of detail. Then, the prediction is multiplied by the albedo ratio of input-a to input-b to produce the relit image.

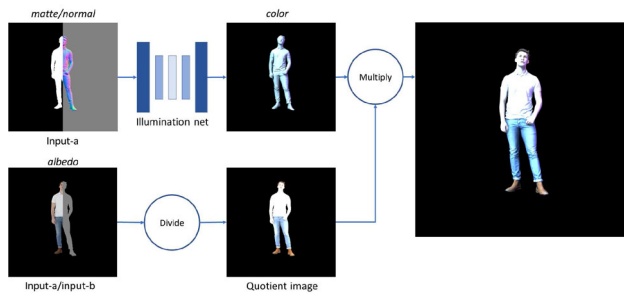


Fig. 5 Lighting transfer module. The illumination net predicts the diffuse reflection component from the alpha matte and surface normals of input-a. By taking the albedo ratio of input-a to input-b, the quotient image is obtained. Finally, the relit foreground image is produced by multiplying the diffuse reflection component by the quotient image

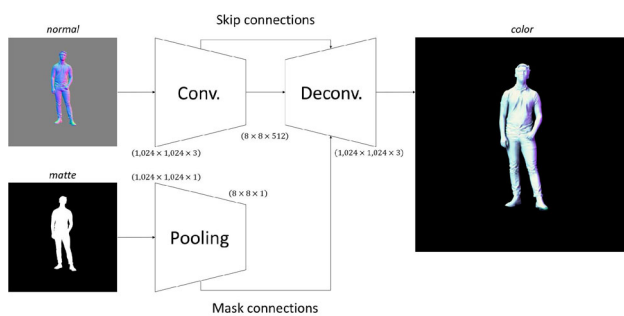


Fig. 6 Architecture of the illumination net. It is based on the U-net and has the mask connections to multiply alpha mattes in the deconvolution layers. The input is a 1024×1024 RGBA image (surface normals + alpha matte), while the output is a 1024×1024 RGB image (diffuse reflection component)

3.4.1 Illumination net

The illumination net predicts the diffuse reflection component from given alpha matte and surface normals. The network is trained using the foreground image, which is the output of the global illumination simulation stage (input-b in Fig. 2).

Network architecture The illumination net is based on U-net [28] and has the mask connections to multiply alpha mattes in the deconvolution layers, as shown in Fig. 6. The input is an RGBA image (surface normals + alpha matte) with the size of 1024×1024 , while the output is an RGB image (diffuse reflection component) with the size of 1024×1024 . The convolution and deconvolution blocks consist of a combination of a 3×3 convolution layer, a batch normalization layer, and a leaky rectified linear unit activation layer. To make the final output HDR, the last part is activated to $[0, +\infty)$ by the rectified linear unit activation layer, except for the batch normalization layer. When implementing, the maximum value was set to the maximum possible value of the single-precision floating-point number.

Training Training was performed on-demand for each scene, and the foreground image output from the global illumination simulation stage (input-b in Fig. 2) was used as the only training data. The network was trained to take the surface normals and alpha matte as the input, and it outputs the color of the diffuse reflection component. Note that the diffuse reflection component is necessary to adjust its HDR values to match the value range $[0, +\infty)$ of the output layer. One of the reasons for the errors in our previous study [19] was the handling manner of these HDR images. If the pixel values are clamped to $[0, 1]$, overexposure occurs in some areas, preventing accurate learning. To improve the generalization performance of the network, the training data were randomly translated in the range of $[-10, 10]$ for the top, bottom, left, and right of the image space, respectively, for data augmentation. The batch size was set to 1, and training was performed for $10 \text{ steps} \times 100 \text{ epochs}$. Mean squared error was used as the loss function, and RMSprop [13] was used as the optimization method. This on-demand training was completed in about 2 minutes.

3.4.2 Relighting

Because the illumination net predicts the diffuse reflection component for the albedo of input-b, it is necessary to retarget the prediction result to the albedo of input-a. We now introduce a quotient image [27], which is defined as the albedo ratio of two objects and is known to be illumination-invariant. If input-a's albedo is denoted as ρ_a and input-b's albedo as ρ_b , the quotient image Q_a for input-a is represented by:

$$Q_a(u, v) = \frac{\rho_a(u, v)}{\rho_b(u, v)},$$

where u, v denote the coordinates in the image space. ρ_a and ρ_b are given from the previous stages, so Q_a is obtained immediately. Finally, by multiplying the predicted diffuse reflection component by Q_a , it is retargeted to input-a.

3.5 Compositing

The composite result is produced by superimposing the foreground image onto the background one. Each pixel value in the composite image C represents the linear interpolation of the pixel value of the foreground image F and the pixel value of the background image B [24]

$$C = \alpha * F + (1 - \alpha) * B,$$

where α ($\in [0, 1]$) denotes the pixel value of foreground alpha matte. Because the composite result is an HDR image, it is finally converted to an SDR image through tone mapping. Figure 7 shows individual composite examples for

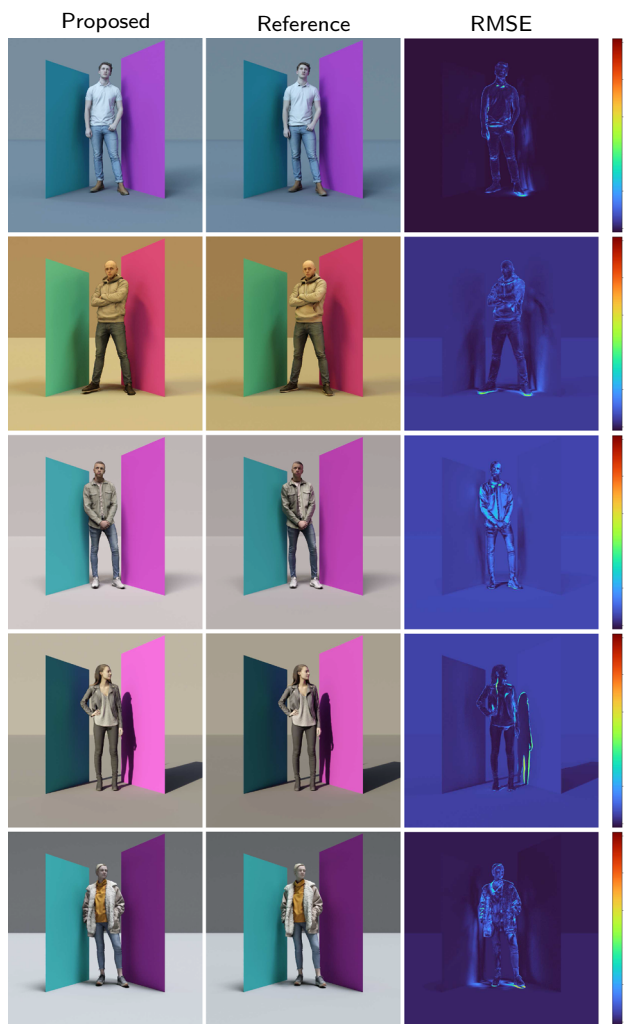


Fig. 7 Individual composite examples for the scenes appearing in Fig. 1. The framework was tested on five different models, each of which was rendered in a different lighting environment. The columns Proposed and Reference show the composite results using the proposed framework and the rendered images of 3D reference models, respectively. The column RMSE visualizes the root mean square error for each pixel between Proposed and Reference with a Turbo colormap [2], which considers visual continuity. The quality of the composite results is comparable to the rendered ones, although an error from tone mapping appears in some of the images

the scenes appearing in Fig. 1. The root mean square error (RMSE) is used for comparison, and the Turbo colormap [2], which maintains visual continuity, is used for visualization. The quality of the composite results is comparable to that of the rendered ones, although an error from tone mapping appears in some of the images. The following link (<https://msrohkw.github.io/Illumination-aware-group-portrait-compositor/>) provides more results.

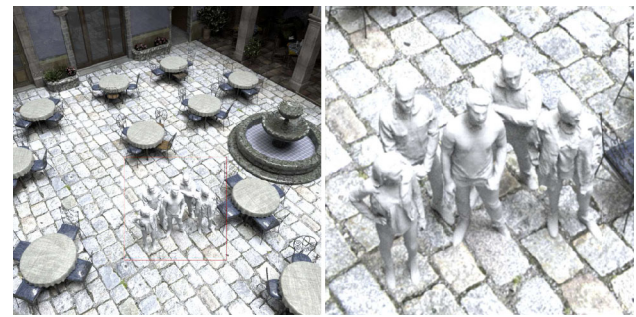




Fig. 8 An overhead view of a group portrait compositing example. The reconstructed shapes are manually placed onto the scene and will be globally illuminated. Note that the output at the global illumination simulation stage is rendered while their relative positioning

4 Group portrait compositing

As a daily life scenario to apply the proposed framework, we present group portrait compositing. One of the roles of group portraits is to give the visual cue that multiple people have shared the same time and space. However, in actual shooting, it is often the case that the space cannot be shared due to the remote collaboration or the time cannot be shared due to the absence of some members. Although, in such cases, convincing group portraits are expected to be produced with compositing techniques. The issue here reduces to visual coherence. When the shooting environment of each person is different, standard composite results are visually incoherent. However, the proposed framework produces group portraits with high visual coherence.

Figure 1 shows the composite results for five individual portraits, each of which was captured under a different lighting environment. The RMSE was used for comparison with the reference image, and a Turbo colormap [2] was used for visualization. The standard composite result is unconvincing in terms of visual coherence, while the composite results produced by the proposed framework compare favorably with the reference images rendered the multiple 3D human models together. In addition, shadows, which cannot be reproduced by common relighting methods, help augment the sense of coexistence in the group portraits. In fact, significant errors can be found between the composite image and the reference image overall, despite some localized errors. Table 1 provides the times required for group portrait compositing. Our method required 4.6 and 3.1 times longer than rendering, respectively, but this seems allowable because the composite results compared favorably with the rendered ones.

Table 1 Times required for group portrait compositing and for rendering are compared. Note that the times at the photometric information estimation stage are not included because the photometric information is assumed to be in this paper

Target	Ours (seconds)						Rendering (seconds)
	Shape reconstruction	Global illumination simulation		Lighting transfer	Compositing	Total	
		Foreground	Background				
	8.7×10^1	2.4×10^1	4.9×10^1	1.1×10^2	2.7×10^{-2}	2.7×10^2	5.9×10^1
	8.7×10^1	2.8×10^1	1.3×10^2	1.1×10^2	2.7×10^{-2}	3.6×10^2	1.2×10^2

In Sect. 3, we discussed individual portrait compositing, so we will now discuss how we extended it to group portrait compositing. Indeed, the extension to group portrait compositing sounds straightforward: a foreground image assuming the composite result can be produced by inputting multiple foreground shapes of people at the global illumination simulation stage, as shown in Fig. 8. At this point, photometric information should be estimated individually at the photometric information estimation stage, be bundled assuming the composite result, and be passed to the lighting transfer stage.

5 Discussions

The current framework has several issues. As described in Sect. 3.1, even though the framework relies on photometric information, it was tested only with rendered images whose actual photometric information was available. In addition, it has the essential issue that supporting various illumination components other than the diffuse reflection is difficult because the materials of the human figures are not reconstructed. Besides, the requirement for manual operation at some stages makes it difficult to realize full automation. In Sects. 5.1 and 5.2, we discuss the availability of the photometric information and the limitations of the material, respectively.

5.1 Availability of photometric information

The accuracy of the photometric information estimation affects the quality of the final results, as shown in Fig. 3. Total relighting [20], which is a deep learning-based state-of-the-art method in portrait relighting, deploys a matting module and relighting module. The matting module estimates the alpha matte of the human, and the relighting module estimates its surface normals and albedo; thus, they can be

utilized for relighting humans. Because OLAT images are required for the training, the model cannot be constructed without special equipment. However, if its pre-trained model is available, the photometric information can be obtained from a single-view RGB image. Because it was reported in [20] that these modules achieved high-quality results, it is technically safe to assume that the desired photometric information can be obtained with these modules. However, these modules suffer from quality degradation due to the albedo imperfections on clothing, so practical verification will be unavoidable.

5.2 Material limitation

To reproduce various illumination components other than diffuse reflection in the composite results, it is required to recover the materials of the human figures at the shape reconstruction stage. Here, it is known that the task of simultaneously recovering texture and material in addition to shape is challenging. PIFu [29], the framework on which PIFuHD is based, can recover shape and texture from a single-view RGB image, but its accuracy is not high enough in terms of content creation, and the material is still not addressed. Therefore, we need to develop a new method to recover the texture and material simultaneously, in addition to the shape, from a single-view image.

6 Conclusion

In this paper, we presented a global illumination-aware compositing framework for full-length human figures and attempted to apply the framework to group portrait compositing. Conventional compositing usually suffers visual incoherence, so the proposed framework can create a composite result with a sense of coexistence that considers optical consistency. In the new era of “remote everything,” provided

that the proposed framework is applied to video conferencing systems, people will be able to experience more unified collaboration. More broadly, it could also contribute to the construction of cyberworlds by leveraging the salient features of the proposed framework, which efficiently generates photorealistic contents.

Acknowledgements This work has been financially supported in part by JSPS KAKENHI Grants-in-Aid for Scientific Research (A) No. 21H04916 and Challenging Research (Pioneering) No. 20K20481.

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Alldieck, T., Pons-Moll, G., Theobalt, C., Magnor, M.: Tex2Shape: Detailed full human body geometry from a single image. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2293–2303 (2019). <https://doi.org/10.1109/ICCV.2019.00238>
2. Anton, M.: Turbo, an improved rainbow colormap for visualization. Blog post on Google AI Blog (2019). <https://ai.googleblog.com/2019/08/turbo-improved-rainbow-colormap-for.html>
3. Autodesk: Aovs - arnold for maya user guide - arnold renderer. Arnold for Maya User Guide (2021). <https://docs.arnoldrenderer.com/display/A5AFMUG/AOVs>
4. Bell, S., Bala, K., Snavely, N.: Intrinsic images in the wild. ACM Trans. Graph. (2014). <https://doi.org/10.1145/2601097.2601206>
5. Blinn, J.F., Newell, M.E.: Texture and reflection in computer generated images. Commun. ACM **19**(10), 542–547 (1976). <https://doi.org/10.1145/360349.360353>
6. Bluff, R., Fields, L., Keller, A., Jones, H., Rose, R.: ILM presents “This is the Way”—the making of Mandalorian. In: ACM SIGGRAPH 2020 Production Sessions, SIGGRAPH ’20. ACM, New York, NY, USA (2020). <https://doi.org/10.1145/3368850.3383439>
7. Cao, X., Shen, Y., Shah, M., Foroosh, H.: Single view compositing with shadows. The Visual Comput. **21**, 639–648 (2005). <https://doi.org/10.1007/s00371-005-0335-x>
8. Chuang, Y.Y., Goldman, D.B., Curless, B., Salesin, D.H., Szeliski, R.: Shadow matting and compositing. ACM Trans. Graph. **22**(3), 494–500 (2003). <https://doi.org/10.1145/882262.882298>
9. Debevec, P.: Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In: Proceedings of the 25th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH ’98, pp. 189–198. ACM, New York, NY, USA (1998). <https://doi.org/10.1145/280814.280864>
10. Debevec, P., Hawkins, T., Tchou, C., Duiker, H.P., Sarokin, W., Sagar, M.: Acquiring the reflectance field of a human face. In: Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH ’00, pp. 145–156. ACM Press/Addison-Wesley Publishing Co., USA (2000). <https://doi.org/10.1145/344779.344855>
11. Du, H., Jin, X.: Object cloning using constrained mean value interpolation. The Visual Comput. **29**, 217–229 (2013). <https://doi.org/10.1007/s00371-012-0722-z>
12. He, T., Xu, Y., Saito, S., Soatto, S., Tung, T.: Arch++: Animation-ready clothed human reconstruction revisited. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 11,046–11,056 (2021)
13. Hinton, G.: Lecture 6d: A separate, adaptive learning rate for each connection. Lecture slides for Neural Networks for Machine Learning (2012). https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf
14. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5967–5976 (2017). <https://doi.org/10.1109/CVPR.2017.632>
15. Jared, S.: The future of work—the good, the challenging & the unknown. Blog post on Microsoft 365 Blog (2020). <https://www.microsoft.com/en-us/microsoft-365/blog/2020/07/08/future-work-good-challenging-unknown/>
16. Kanamori, Y., Endo, Y.: Relighting humans: Occlusion-aware inverse rendering for full-body human images. ACM Trans. Graph. (2018). <https://doi.org/10.1145/3272127.3275104>
17. Kwatra, V., Schödl, A., Essa, I., Turk, G., Bobick, A.: Graphcut textures: Image and video synthesis using graph cuts. In: ACM SIGGRAPH 2003 Papers, SIGGRAPH ’03, pp. 277–286. ACM, New York, NY, USA (2003). <https://doi.org/10.1145/1201775.882264>
18. Li, Q., Yin, W.: Image-based face illumination transferring using logarithmic total. The Visual Computer **26**, 41–49 (2010). <https://doi.org/10.1007/s00371-009-0375-8>
19. Ohkawara, M., Fujishiro, I.: Illumination-aware digital image compositing for full-length human figures. In: 2021 International Conference on Cyberworlds (CW), pp. 17–24 (2021). <https://doi.org/10.1109/CW52790.2021.00011>
20. Pandey, R., Escolano, S.O., Legendre, C., Häne, C., Bouaziz, S., Rhemann, C., Debevec, P., Fanello, S.: Total relighting: Learning to relight portraits for background replacement. ACM Trans. Graph. (2021). <https://doi.org/10.1145/3450626.3459872>
21. Pérez, P., Gangnet, M., Blake, A.: Poisson image editing. ACM Trans. Graph. **22**(3), 313–318 (2003). <https://doi.org/10.1145/882262.882269>
22. Philip, J., Gharbi, M., Zhou, T., Efros, A.A., Drettakis, G.: Multi-view relighting using a geometry-aware network. ACM Trans. Graph. (2019). <https://doi.org/10.1145/3306346.3323013>
23. Pitie, F., Kokaram, A., Dahyot, R.: N-dimensional probability density function transfer and its application to color transfer. In: Tenth IEEE International Conference on Computer Vision (ICCV’05), vol. 2, pp. 1434–1439 (2005). <https://doi.org/10.1109/ICCV.2005.166>
24. Porter, T., Duff, T.: Compositing digital images. SIGGRAPH Comput. Graph. **18**(3), 253–259 (1984). <https://doi.org/10.1145/964965.808606>
25. Qin, X., Zhang, Z., Huang, C., Dehghan, M., Zaiane, O.R., Jagersand, M.: U²-net: Going deeper with nested u-structure for salient object detection. Pattern Recogn. (2020). <https://doi.org/10.1016/j.patcog.2020.107404>
26. Reinhard, E., Ashikhmin, M., Gooch, B., Shirley, P.: Color transfer between images. IEEE Comput. Graph. Appl. **21**(5), 34–41 (2001). <https://doi.org/10.1109/38.946629>
27. Riklin-Raviv, T., Shashua, A.: The quotient image: Class based recognition and synthesis under varying illumination conditions. In: Proceedings 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149), vol. 2, pp. 566–571 (1999). <https://doi.org/10.1109/CVPR.1999.784968>
28. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional networks for biomedical image segmentation. In: N. Navab, J. Hornegger, W.M. Wells, A.F. Frangi (eds.) Proceedings of the Medical Image Computing and Computer-Assisted Intervention—MICCAI 2015, pp. 234–241 (2015). https://doi.org/10.1007/978-3-319-24574-4_28
29. Saito, S., Huang, Z., Natsume, R., Morishima, S., Li, H., Kanazawa, A.: PIFu: Pixel-aligned implicit function for high-resolution

- clothed human digitization. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 2304–2314 (2019). <https://doi.org/10.1109/ICCV.2019.00239>
30. Saito, S., Simon, T., Saragih, J., Joo, H.: PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 81–90 (2020). <https://doi.org/10.1109/CVPR42600.2020.00016>
 31. Shu, Z., Hadap, S., Shechtman, E., Sunkavalli, K., Paris, S., Samaras, D.: Portrait lighting transfer using a mass transport approach. *ACM Trans. Graph.* (2017). <https://doi.org/10.1145/3095816>
 32. Sun, T., Barron, J.T., Tsai, Y.T., Xu, Z., Yu, X., Fyffe, G., Rhemann, C., Busch, J., Debevec, P., Ramamoorthi, R.: Single image portrait relighting. *ACM Trans. Graph.* (2019). <https://doi.org/10.1145/3306346.3323008>
 33. Susanna, R.: Video fatigue and a late-night host with no audience inspire a new way to help people feel together, remotely. Story on Microsoft Innovation Stories (2020). <https://news.microsoft.com/innovation-stories/microsoft-teams-together-mode/>
 34. Tingbo, H., Tyler, M.: Background features in Google Meet, powered by Web ML. Blog post on Google AI Blog (2020). <https://ai.googleblog.com/2020/10/background-features-in-google-meet.html>
 35. Tsai, Y.H., Shen, X., Lin, Z., Sunkavalli, K., Lu, X., Yang, M.H.: Deep image harmonization. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2799–2807 (2017). <https://doi.org/10.1109/CVPR.2017.299>
 36. Wang, T.C., Liu, M.Y., Zhu, J.Y., Tao, A., Kautz, J., Catanzaro, B.: High-resolution image synthesis and semantic manipulation with conditional gans. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 8798–8807 (2018). <https://doi.org/10.1109/CVPR.2018.00917>
 37. Wang, Y., Liu, A., Tucker, R., Wu, J., Curless, B.L., Seitz, S.M., Snavely, N.: Repopulating street scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5110–5119 (2021)
 38. Xu, D., Li, Z., Cao, Q.: Object-based illumination transferring and rendering for applications of mixed reality. *The Visual Comput.* (2021). <https://doi.org/10.1007/s00371-021-02292-2>
 39. Yun-Ta, T., Rohit, P.: Portrait light: Enhancing portrait lighting with machine learning. Blog post on Google AI Blog (2020). <https://ai.googleblog.com/2020/12/portrait-light-enhancing-portrait.html>
 40. Zheng, Z., Yu, T., Wei, Y., Dai, Q., Liu, Y.: DeepHuman: 3D human reconstruction from a single image. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), pp. 7738–7748 (2019). <https://doi.org/10.1109/ICCV.2019.00783>



Masaru Ohkawara received his B.E and M. E. in information and computer science in 2019 and 2021 from Keio University, Japan, respectively. He is currently working toward his Ph. D. degree at the Graduate School of Science and Technology, Keio University. His main research interests include visually coherent image synthesis.



Issei Fujishiro received his B.E. and M.E. in information sciences and electronics in 1983 and 1985 from University of Tsukuba, and his Doctor of Science in information sciences from the University of Tokyo in 1988. He is currently a Professor at Department of Information and Computer Science, Faculty of Science and Technology, Keio University, Japan, with an adjunct professorship at School of Computer Science, Hangzhou Dianzi University, China. His research interests

include modeling paradigms and shape representations, applied visualization design and lifecycle management, and smart ambient media underpinned by perceptive psychology.