Eurographics Symposium on Rendering (2024)
E. Garces and E. Haines (Editors)

Learning Self-Shadowing for Clothed Human Bodies

Farshad Einabadi, Jean-Yves Guillemaut and Adrian Hilton

Centre for Vision, Speech and Signal Processing (CVSSP), University of Surrey, Guildford, England
{f.einabadi, j.guillemaut, a.hilton}@surrey.ac.uk

Abstract

This paper proposes to learn self-shadowing on full-body, clothed human postures from monocular colour image input, by supervising a deep neural model. The proposed approach implicitly learns the articulated body shape in order to generate self-shadow maps without seeking to reconstruct explicitly or estimate parametric 3D body geometry. Furthermore, it is generalisable to different people without per-subject pre-training, and has fast inference timings. The proposed neural model is trained on self-shadow maps rendered from 3D scans of real people for various light directions. Inference of shadow maps for a given illumination is performed from only 2D image input. Quantitative and qualitative experiments demonstrate comparable results to the state of the art whilst being monocular and achieving a considerably faster inference time. We provide ablations of our methodology and further show how the inferred self-shadow maps can benefit monocular full-body human relighting.

CCS Concepts

• ***Computing methodologies*** → *Image-based rendering;* ***Visibility; Neural networks;***

1. Introduction

Modelling and rendering self-shadowing, or equivalently self-visibility, as well as cast shadows, are extensively studied in computer graphics literature through shadow mapping [SWP11], real-time soft shadows [HLHS03], shadow volumes [LWGM04], and lighting-independent ambient occlusion shading effects [RBA09]. Ray-tracing methods trace (many) shadow rays from every surface point to calculate lighting visibilities and contributions [PJH16].

However, all of these well-known techniques require the 3D scene geometry, in some form of representation, to be able to render shadowing effects. Recently, there are some efforts to learn to render these effects without the knowledge of geometry for various solid [SZP*23] or articulated object categories such as humans [ZLWY23, CL22, SCHG23, EGH23]. This is of particular interest for realistic lighting effects in mixed reality scenes containing foreground actors or presenters, which are captured from monocular views in front of chroma key backgrounds in studios.

Recent methods that model clothed human self-shadowing in/for the relighting process fall into two categories. The first class of approaches require first estimating the corresponding 3D body geometry explicitly, in the form of triangle meshes [JYG*22], or point clouds [ZLWY23], or implicitly through parametric human models [CL22], from monocular or multiple views, which, in practice, is a computationally expensive task. The second class of methods are based on spherical harmonics (SH) bases, in image space, to represent lighting, and light transport (visibilities, surface normals, and potentially materials), which, although fast, suffer from modelling high-frequency shadowing effects [JYG*22, NRH03].

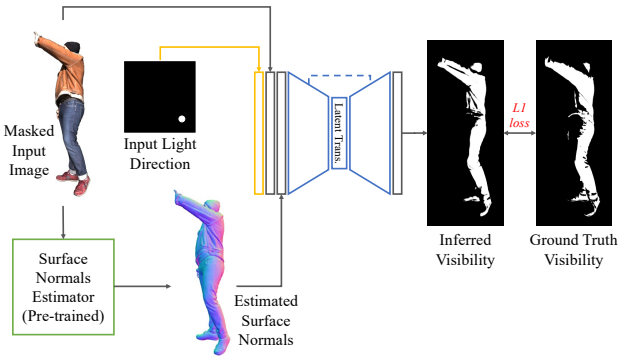

Figure 1: *Overview of our method to infer a self-shadow mask (visibility) given an input image cut-out, and the desired light direction. Surface normals are generated using an available, pre-trained model and fed to the network as an auxiliary signal.*

In this paper, to address both of the above issues, we propose to learn self-shadow (or visibility) *binary maps* for a given (distant) light direction on articulated human postures without explicit estimation, or having the prior of 3D body geometry (see Figure 1). More specifically, the self-shadow maps are defined as the binary pixel values in the camera image space, corresponding to occlusion of points on the object, here the subject's body, from an incoming light direction view point.

DOI: 10.2312/sr.20241159

We propose a deep convolutional encoder-decoder architecture to learn the transformation from an input colour image cut-out, and a light direction to the corresponding self-shadow map. Surface normals are generated using a pre-trained network – extracted from PIFuHD [SSSJ20] – and fed to the model as an auxiliary signal. We train the proposed model on ray-traced self-shadow maps of 3D scanned models taken from 3D People [3DP] for various light directions and with data augmentation. We use the remaining models from 3D People that have not been used for training to evaluate the proposed method's generalisation to unseen people. We provide ablations of our approach with regard to the choice of model's latent space transformation, and the input image saturation levels. We further evaluate our method on the subjects of the People Snapshot Dataset [AMX*18] using the pre-trained, *person-specific* models provided by Relighting4D [CL22]. Compared to Relighting4D, we demonstrate improvements in the inference speed (about 1 to 2 orders of magnitude faster), and generalisation, whilst our approach does not require per-person pre-training with many frames. Finally, we show experimental evaluation how the generated self-shadow maps as a pre-trained auxiliary signal can improve estimated diffuse shadings in the context of monocular full-body human relighting.

In summary, our main contributions are:

- A fast, generalisable, neural model to infer self-shadow maps for clothed human bodies, from monocular input colour image cut-outs, without the prior knowledge of or the requirement to estimate explicit/implicit 3D human body geometry; and
- Demonstration of the use of estimated self-shadows for improved relighting of people in monocular setup without a requirement for 3D shape estimation.

2. Related Work

In this section, we focus on the work most related to learning visibility for human relighting. For the advances in neural methods for rendering and (re)lighting, refer to the recent surveys by Tewari et al. [TTM*22], and Einabadi et al. [EGH21], respectively.

Explicit Geometry. Zheng et al. [ZLWY23] learn visibility fields for a uniform, discrete set of 64 directions from multiple RGB-D views, which are then used to render shading images with a physically-based process. However, multiple calibrated views including depth values are needed to build a prior point cloud of the subject in the first stage. Similarly, Chen and Liu [CL22] is geometrically conditioned on a parametric human model, SMPL [LMR*15] or its variants, estimated from input video in an iterative, time-consuming manner. Here, estimated vertex-dependent latent features are used to infer visibility maps for a discrete light source direction, among 32×16, i.e. 512 possibilities, via a fully-connected, multilayer submodule – which is naturally slow due to per (chunk of) pixel(s) inference compared to our proposed convolutional model. Both of these methods [ZLWY23, CL22] are person-specific and not generalisable to new people after training. Ji et al. [JYG*22] estimate explicit 3D meshes using PIFuHD [SSSJ20] from monocular input image, which are then fed into a path tracer with a neural refinement module to infer the final shading. Both the 3D mesh estimation and its rendering via a path-tracer are computationally expensive.

Note that Iqbal et al. [ICN*23] and Sun et al. [SCHG23] explicitly estimate 3D geometry (respectively through a SMPL model, and 3D canonical volumes), but do not model visibility (or equivalently occlusions), i.e. only the subsequently derived surface normals are employed in the renderer module to relight humans. Therefore, relit images unavoidably do not demonstrate visibility-related self-shadowing effects on the body. Also, the training of Sun et al. [SCHG23] is person-specific.

In comparison, our proposed approach does not require estimating explicit parametric or non-parametric 3D geometry of the human body for modelling visibility – which are yet to be rendered by neural or physically-based renderers – and is therefore faster in inference. Also, the training procedure of our model is not person-specific and generalises to different people and clothing.

Image-based Modelling of Light Transport. The seminal relighting work of Kanamori and Endo [KE18] models clothed human body visibility with surface normals inseparably included (*baked-in*) represented by second-order spherical harmonics, from monocular input colour images. Relit images are rendered for image-space by the dot product of the SH coefficients of light transport, and target lighting. Tajima et al. [TKE21] build upon the previous work [KE18] by adding a second *photo-domain adaptation* step for enhanced realism, and improvements to remove diffuse albedo-light colour inference ambiguity. Lagunas et al. [LSY*21] enhance the work of Kanamori and Endo [KE18] by lifting the assumption of Lambertian material, by training a similar model architecture on synthetically generated images of human models with specular reflectance material properties. This category of methods [LSY*21, TKE21, KE18] are generalisable to different people.

Our method, on the contrary, models visibility separately (not baked-in), and further improves on the inference quality by not being limited to spherical harmonic representation – this limitation is specifically reported and evaluated by Ji et al. [JYG*22] as [LSY*21, TKE21, KE18] are unable "to model high-frequency shadows due to reliance on spherical harmonics representation of lighting." Also, it is noteworthy that a low order SH projection inherently cannot precisely model and localise a directional illumination signal (see Appendix A).

3. Methodology

In this section, first we briefly describe the proposed model. Then, we present the training data generation process, and the training and implementation details.

Model. The proposed encoder-decoder architecture with residual blocks [HZRS16] in the latent space is based on U-Net [RFB15] and is depicted in Figure 1 and Table 1. U-Net-like architectures have successfully been used in image-to-image transformation tasks in general [IZZE17], and recently for estimating shadowing effects [SLZ*22, LLZ*20].

Note that skip connections exist between each 5 corresponding down- (DS) and up-sampling (US) blocks. Each DS/US block contains multiple sets of convolution and instance normalisation layers, followed by a last bilinear upsampling layer in the US block, where the last convolution layer in the DS block has stride 2. The model has 45.8 million parameters.

Table 1: *The convolutional model architecture*

Module	Layer	Kernel	Resample	Output
	Input	-		$512 \times 512 \times 9$
Encoder	DS$\times$5	3×3	Stride 2	$16 \times 16 \times 512$
Latent T.	ResB $\times$2	7×7	-	$16 \times 16 \times 512$
Decoder	UL$\times$5	3×3	Upsample $\times$2	$512 \times 512 \times 1$
	tanh	-	-	$512 \times 512 \times 1$

The inputs to the model are a masked RGB image, auxiliary surface normals estimated by the pre-trained PIFuHD [SSSJ20] *frontal* surface normal generator, and the target light direction (in unit vector representation), and so in total 9 channels. The output is the inferred 1-channel binary self-shadow map. In Section 4.2, we ablate our model with regard to the choice of the latent space transformation.

Data Generation. We take 150 3D scanned models from the 3D People dataset [3DP], 120 for training (of which 10% for validation), and 30 reserved for generalisation evaluations. Each model is rendered from 9 viewpoints distributed around the subject, at a distance of 175 cm and a height of 90 cm from the floor. The camera's field of view is 70 degrees and the principal axis is parallel to the floor. All images are rendered by pbrt-v3 [PJH16] with the resolution of 512×512, and 32 samples per pixels. There are in total about 62k training, 7k validation, and 17k generalisation self-shadow maps in the dataset.

The choice of the camera parameters for generating training and test images is to motivate *normalised* appearance of the subject in the images with regard to scale and position. This, in principle, does not affect the generalisation capability of the model regarding the subject's scale and position, but does so for different camera angles due to various elevations. Our generated training data is, e.g. similar to the training data used in the previous work PIFuHD [SSSJ20] – with weak-perspective camera assumption – where the camera moves around the subject at a fixed elevation.

Each dataset entry contains a binary mask, surface normals, and *diffuse* albedo images, and a set of 64 self-shadow maps and path-traced *diffuse* shadings, corresponding to 64 uniformly sampled, discrete directions on the (frontal) hemisphere facing the camera. The input colour images for the training and inference are randomly lit, i.e. they are random mixture of the images rendered in image space based on the dot product of surface normals, and a light direction, then multiplied by a self-shadow map, and the albedo – this is equivalent to path-traced images with a path length of 1, i.e. only direct lighting. Furthermore, a random overall light intensity is also applied to the colour images. The ground truth binary self-shadow maps are rendered by tracing shadow rays from the camera ray-body intersection points in the direction of incoming light to detect self-intersection (self-occlusion). Refer to Figures 1 and 3 for samples.

Training and Implementation Details. The model is implemented in PyTorch and trained using Adam optimiser with L1 reconstruction loss ($1/n \sum_{i=1}^{n} |I_i^{pred} - I_i^{gt}|$), mini-batch size of 1, and the learning rate of 5e$-$6 for 31 epochs on a GeForce RTX 2080 for about 25 hours. The inference time is about 14 ms for one sample for our model, and about 8 ms for the surface normals generator extracted from PIFuHD [SSSJ20]. In our setup, for burst evaluations, these

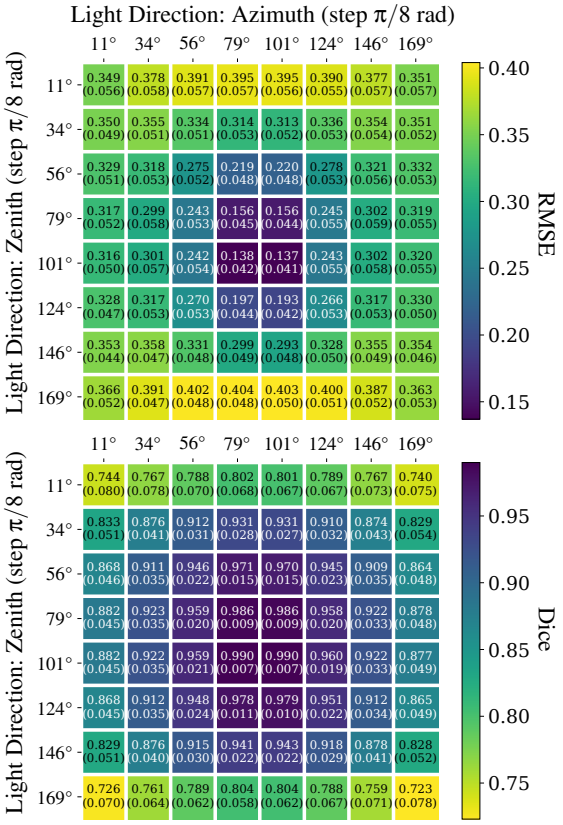

Figure 2: *Self-shadow map inference error heatmaps, average (standard deviations) for the 64 frontal discrete light directions*

timings converge to 27 ms and 70 ms, respectively. For comparison, it is noteworthy that 3D geometry estimation from a monocular input image by PIFuHD [SSSJ20] takes about 12 seconds in the same setup, which still needs to be rendered by, e.g. a path tracer.

4. Experiments

In this section, we first evaluate the proposed model on the rendered self-shadow map generalisation set of Section 3 and further provide the corresponding methodology ablations. Then we compare our method to Relighting4D [CL22], which requires explicit 3D human shape reconstruction from (monocular) multiple frames, for a number of pre-trained real subjects of the People Snapshot Dataset [AMX*18]. Finally, we demonstrate in an experiment how self-shadow maps can help improve monocular estimation of diffuse shadings regarding self-shadowing effects.

4.1. Evaluation on the Generalisation Set of Section 3

Metrics. We report Root Mean Square Error (RMSE), Peak Signal-to-Noise Ratio (PSNR), and Dice (F1 Score) metrics measured on binary self-shadow map images. All metrics are calculated in the mask region. Since the maximum pixel value is 1 for binary self-shadow maps, PSNR (in dB) is related to RMSE by $-20 \log_{10}(\text{RMSE})$.

Baseline. As some visibility information exists in surface normals,

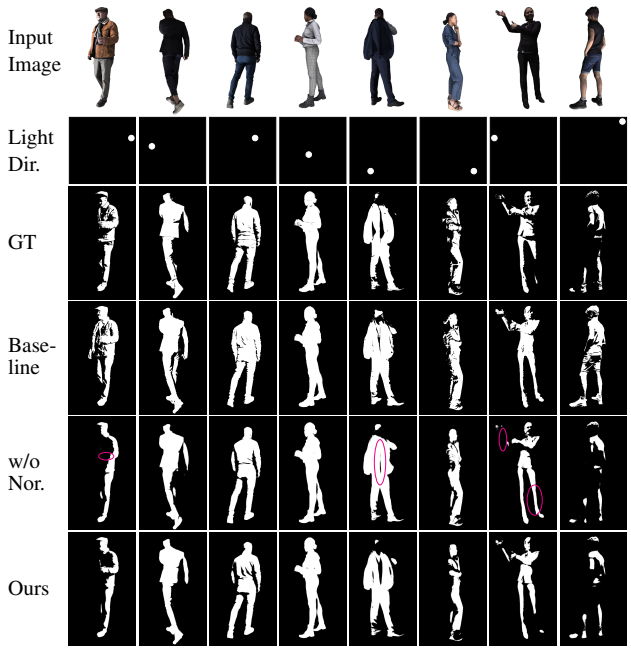

Figure 3: *Inferred self-shadows maps for representative samples of the rendered generalisation set. The distant light directions in row 2 correspond to the centres of the blobs, the area of blobs are for visualisation purposes only.*

Table 2: *Quantitative comparison to the baseline. Average (standard deviation) calculated on the rendered generalisation set*

Variant	RMSE ↓	PSNR ↑	Dice ↑
Baseline	0.432 (0.056)	7.290	0.834 (0.049)
Ours	**0.313 (0.051)**	**10.089**	**0.881 (0.041)**

a fast, naive baseline is considered for the evaluations where the output binary self-shadow maps are generated based on the *sign* of the dot product of the surface normal of a point and the incoming light direction.

Results and Discussion. Figures 2 and 3 respectively show quantitative and qualitative results and demonstrate that our method is able to generalise to unseen images of people. The inference error is lower for light directions almost parallel to the camera view, where self-shadowing is minimal, and gradually increases towards the periphery, as depicted by both heatmaps.

Figure 3 demonstrates that the baseline (row 4) has visibly no notion of self-shadowing effects resulting in large errors in the self-shadow masks. Table 2 provides the quantitative results compared to the baseline.

Furthermore, using the additional pre-trained surface normals estimator as an auxiliary signal (Figure 3 row 6) helps improve some depth ambiguity and clothing details artefacts compared to without (Figure 3 row 5, some artefacts highlighted).

4.2. Ablations

Surface Normals. Table 3 shows the calculated metrics for three variants of employed surface normals in the training and inference phases. The results demonstrate that using the extracted surface normals generator of PIFuHD [SSSJ20] as an auxiliary signal to the model for both training and inference phases (row 3) improves the performance compared to not using them at all (row 1). Also, the performance of the mixed usage of ground truth and PIFuHD surface normals (row 2) is lower than the without case.

Table 3: *Ablations with regard to surface normals. Average (standard deviation) calculated on the rendered generalisation set*

Variant	Training	Inference	RMSE ↓	Dice ↑
w/o Nor.	–	–	0.334 (0.054)	0.867 (0.045)
Mixed	GT	PIFuHD	0.352 (0.053)	0.862 (0.044)
Ours	PIFuHD	PIFuHD	**0.313 (0.051)**	**0.881 (0.041)**

Table 4: *Ablation of latent space transformation. Average (standard deviation) calculated on the rendered generalisation set*

Variant	Infer. (ms) ↓	Params. (mil.) ↓	RMSE ↓	Dice ↑
ResB 3	**11.9**	**24.8**	0.318 (0.053)	0.879 (0.042)
ResB 7	14.2	45.8	**0.313 (0.051)**	**0.881 (0.041)**
ResB 11	15.7	83.5	0.315 (0.052)	0.879 (0.041)
Self-Att.	15.9	39.0	0.322 (0.053)	0.876 (0.042)

Latent Space Transformation. We further evaluated the performance of the model with regard to the latent space transformation (depicted in Figure 1). Table 4 shows the effects of increasing the convolution kernel sizes of the latent residual blocks [HZRS16] from 3 to 11 (inspired from related neural shadowing works in the literature [EGH23, ZLW19]), compared to employing a self-attention (Self-Att.) module. This ablation is to take into consideration the global nature of shadow transformation in the image space. The results show residual blocks of various kernel sizes are performing comparably to the self-attention module in terms of computational costs and performance.

The self-attention module employed here is a stack of 4 identical multi-headed attention (MHA) layers [VSP*17] with the corresponding normalisation and linear projections, having the same query, key and values as input.

Input Image Saturation. We ablate the performance of the model versus the input intensity levels. Figure 4 presents Dice metric values for various input intensities, ranging from very dark, i.e. a black

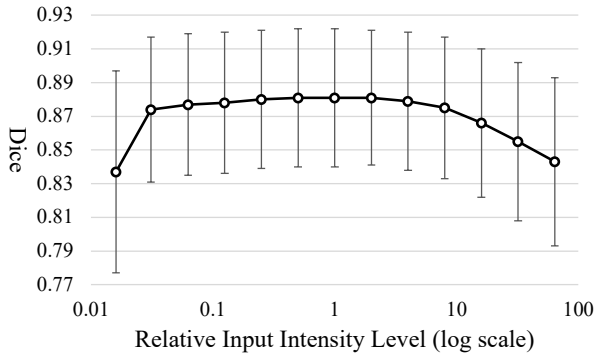

Figure 4: *Input saturation level ablation. Average (standard deviations) for the 64 frontal discrete light directions*

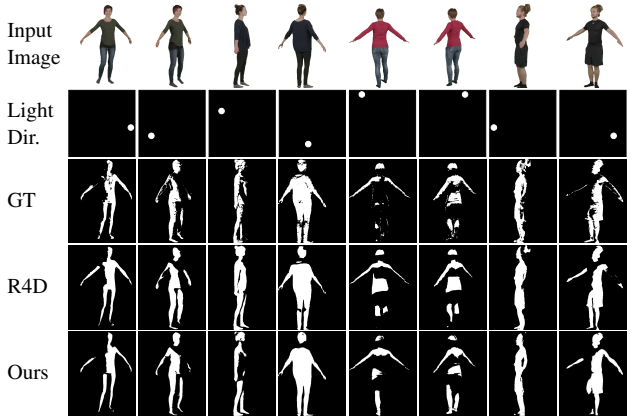

Figure 5: *Qualitative comparisons to Relighting4D [CL22]*

image (intensity levels 0.01 and below), to very bright, i.e. a mask image (intensity levels 10 and above). We observe that the performance on the extreme sides of the spectrum, i.e., lack of a meaningful input colour images, is comparable to the surface normals-based baseline mentioned in Section 4.1. Black input images cause failure in PIFuHD surface normal estimator, whereas PIFuHD is still able to estimate smooth surface normals, with depth ambiguity artefacts, for bright input images, i.e. almost binary masks. In other cases, the model is however resistant to input saturation levels and maintains its performance.

4.3. Comparison to Relighting4D [CL22]

Figure 5 and Table 5 compare our method to the state-of-the-art neural inverse rendering of human appearance from videos, Relighting4D [CL22], on pre-trained, person-specific neural representations of 4 *real* subjects in the People Snapshot Dataset [AMX*18], 1 male and 3 females, on the above 64 frontal light directions, and 11 distinctive frames per subject.

The inferred self-shadow maps of Relighting4D [CL22] are not binary. In our comparisons, we hence first threshold them using Otsu's method [Ots79] which maximises between-class variance in the corresponding grey-level histograms. Results show that our *monocular* method has comparable metric values to Relighting4D, without having seen the images beforehand nor the need for prior person-specific training of each subject.

In terms of inference speed, Relighting4D is conditioned on the SMPL [LMR*15] geometry model, which requires an iterative process to estimate the corresponding parameters, which might take up to 1 or 2 seconds. Furthermore, latent features of a position on a neural field are fed into separate, fully-connected layers for each desired element, e.g., occlusion map, albedo, etc., which operate per (chunk of) pixel(s) and are slower compared to their convolutional counterparts. In our setup, estimating the 512×512 occlusion maps for a set of fixed light directions took on average 8.5 s for f3c in Table 5, or similar for the other models.

Comparison to Other Methods. Lagunas et al. [LSY*21], Tajima et al. [TKE21], Kanamori and Endo [KE18] do not explicitly model self-shadowing, but rather implicitly baked in light transport. In addition, the methods [LSY*21, TKE21, KE18] use low order spherical harmonics representation of the lighting, which cannot precisely reconstruct directional lights for our one-light-at-a-time (OLAT) relighting (Appendix A). Iqbal et al. [ICN*23] and Sun et al. [SCHG23] do not model self-shadowing at this stage. Zheng et al. [ZLWY23] is not monocular and requires depth values. Ji et al. [JYG*22] estimate proxy 3D mesh geometry to ray-trace self-shadowing, which we are avoiding.

Table 5: *Quantitative comparisons to Relighting4D [CL22]. Average (standard deviation) on 11 distinctive frames per subject*

	RMSE ↓		Dice ↑	
Subject	Relighting4D	Ours	Relighting4D	Ours
f1c	**0.404 (0.086)**	0.423 (0.088)	**0.849 (0.095)**	0.815 (0.114)
f3c	**0.401 (0.093)**	0.434 (0.096)	**0.854 (0.095)**	0.812 (0.120)
f4c	**0.403 (0.087)**	0.425 (0.094)	**0.848 (0.096)**	0.812 (0.124)
m5s	**0.429 (0.102)**	0.466 (0.096)	**0.824 (0.119)**	0.777 (0.136)

4.4. Self-shadow maps for monocular OLAT relighting (diffuse shading)

In this section, we employ the ResB 7 variant of the model described in Section 3 (with the final *linear* activation) to learn the diffuse shading with high-frequency self-shadowing effects in the context of fast, monocular, OLAT-based, full-body human relighting. The training is performed in linear space (without tone mapping), and the output of the network is normalised, without loss of generality, to have maximum shading value of 1. As such, given intensity of the light source, it can be multiplied to the network's output for relighting purposes. It is also noteworthy that the model learns the difference to a base shading image rendered using the estimated surface normals, and the estimated visibility maps, if either or both available as auxiliary signal(s) in the ablations.

For the quantitative results, RMSE (masked) and Structural Similarity Index Measure (SSIM) are reported in Table 6 compared to ground truth *path-traced* diffuse shadings. SSIM is calculated for the bounding box region of the mask. Figure 6 demonstrates the corresponding qualitative results for our model and its ablations. This experiment shows the benefits of using self-shadow maps as an additional input signal (from our pre-trained module) for the aforementioned relighting problem formulation, whilst the absolute error heatmaps of rows 7 and 8 show visibility-related artefacts respectively for the variant which uses the baseline visibility of Section 4.1, and the without visibility variant. Also, the without visibility variant sometimes suffers from checkered artifacts, e.g., in the second and eighth columns, also reported by Ji et al. [JYG*22] to be "due to the limitations of the PIFuHD network's generation capacity and memory space."

Moving Light Source. Additional material contains videos of subjects lit with a moving light source – rotating in the frontal hemisphere from right to left with various elevations – and the corresponding ground truth and our self-shadow maps, as well as its application in monocular diffuse shading. The results suggests our approach does not suffer from artefacts related to temporal changes in the light source direction.

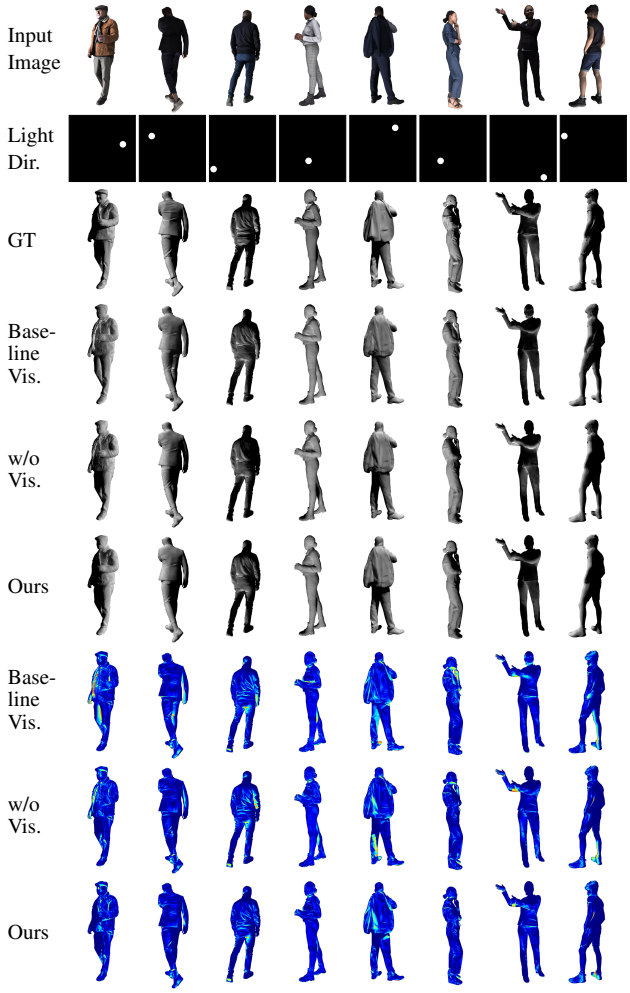

Figure 6: *Inferred diffuse shadings for representative samples of the rendered generalisation set*

4.5. Limitations

Our approach is dependent on the precision of the pre-trained surface normals estimator module (here extracted from PIFuHD [SSSJ20]) and can have lower visual fidelity of details compared to the ground truth, for both self-shadow maps and the diffuse shadings. Similar to the previous work PIFuHD [SSSJ20], our dataset is generated with the camera rotating around the subjects at a fixed height, which might therefore limit the generalisation capability of our model for test images captured with substantially different camera elevations. Furthermore, although we have comparable results to the methods using explicit 3D geometry, our approach can suffer from depth ambiguities due to the 2D monocular input. Figure 7 shows examples of this phenomenon for hand shadows.

5. Conclusion

We presented a fast, generalisable method for estimating self-shadow maps on clothed human bodies and evaluated its performance qualitatively and quantitatively against unseen full-body images. Compared to the state of the art, we demonstrated comparable self-shadow map generation accuracy, whilst being monocular, and reducing inference times by about 1 to 2 orders of magnitude. We further show how self-shadowing as pre-trained auxiliary signal can improve monocular, OLAT-based, full-body human relighting, demonstrating high-frequency self-shadowing effects. Avenues for further research are, for example, modelling temporal coherence in videos, evaluation of other potential training losses, as well as evaluating other shadow image representations such as that of Griffiths et al. [GRP22] to potentially further facilitate the learning phase.

Table 6: *Ablation of inferred diffuse shadings with regard to auxiliary visibility and surface normal signals. Average (standard deviation) calculated on the rendered generalisation set*

Variant	Visibility	Surface Normals	RMSE ↓	SSIM ↑
Baseline	Baseline of Sec. 3	PIFuHD	0.153 (0.024)	0.702 (0.062)
Vis.				
w/o	–	–	0.140 (0.025)	0.746 (0.057)
w/o Vis.	–	PIFuHD	0.139 (0.023)	0.734 (0.058)
w/o Nor.	Ours	–	0.137 (0.023)	0.749 (0.056)
Ours	Ours	PIFuHD	**0.135 (0.024)**	**0.751 (0.056)**

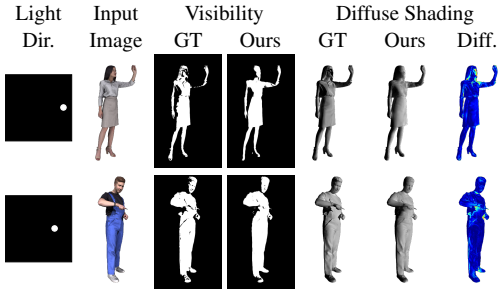

Figure 7: *Error in estimating the self-shadow maps and its impact on diffuse shadings. Missing hand shadows visible in the error heatmaps.*

Acknowledgements

This research was supported by UKRI EPSRC BBC Prosperity Partnership AI4ME: Future Personalised Object-Based Media Experiences Delivered at Scale Anywhere EP/V038087/1.

References

[3DP] 3DPEOPLE UG: 3DPeople. 3dpeople.com. 2, 3

[AMX*18] ALLDIECK T., MAGNOR M., XU W., THEOBALT C., PONS-MOLL G.: Video based reconstruction of 3D people models. In *CVPR* (2018), pp. 8387–8397. 2, 3, 5

[CL22] CHEN Z., LIU Z.: Relighting4D: Neural relightable human from videos. In *ECCV* (2022), pp. 606–623. 1, 2, 3, 5

[EGH21] EINABADI F., GUILLEMAUT J.-Y., HILTON A.: Deep neural models for illumination estimation and relighting: A survey. *Comput. Graph. Forum 40*, 6 (2021), 315–331. 2

[EGH23] EINABADI F., GUILLEMAUT J.-Y., HILTON A.: Learning projective shadow textures for neural rendering of human cast shadows from silhouettes. In *EGSR* (2023), pp. 63–75. 1, 4

[GRP22] GRIFFITHS D., RITSCHEL T., PHILIP J.: OutCast: Outdoor single-image relighting with cast shadows. *Comput. Graph. Forum 41*, 2 (2022), 179–193. 6

[HLHS03] HASENFRATZ J. M., LAPIERRE M., HOLZSCHUCH N., SILLION F.: A survey of real-time soft shadows algorithms. *Comput. Graph. Forum 22*, 4 (2003), 753–774. 1

[HZRS16] HE K., ZHANG X., REN S., SUN J.: Deep residual learning for image recognition. In *CVPR* (2016), pp. 770–778. 2, 4

[ICN*23] IQBAL U., CALISKAN A., NAGANO K., KHAMIS S., MOLCHANOV P., KAUTZ J.: RANA: Relightable articulated neural avatars. In *ICCV* (2023), pp. 23142–23153. 2, 5

[IZZE17] ISOLA P., ZHU J.-Y., ZHOU T., EFROS A. A.: Image-to-image translation with conditional adversarial networks. In *CVPR* (2017), pp. 1125–1134. 2

[JYG*22] JI C., YU T., GUO K., LIU J., LIU Y.: Geometry-aware single-image full-body human relighting. In *ECCV* (2022). 1, 2, 5

[KE18] KANAMORI Y., ENDO Y.: Relighting humans: Occlusion-aware inverse rendering for full-body human images. *ACM Trans. Graph. 37*, 6 (2018). 2, 5

[LLZ*20] LIU D., LONG C., ZHANG H., YU H., DONG X., XIAO C.: ARShadowGAN: Shadow generative adversarial network for augmented reality in single light scenes. In *CVPR* (2020). 2

[LMR*15] LOPER M., MAHMOOD N., ROMERO J., PONS-MOLL G., BLACK M. J.: SMPL: A skinned multi-person linear model. *ACM Trans. Graph. 34*, 6 (2015). 2, 5

[LSY*21] LAGUNAS M., SUN X., YANG J., VILLEGAS R., ZHANG J., SHU Z., MASIA B., GUTIERREZ D.: Single-image Full-body Human Relighting. In *EGSR (DL)* (2021), pp. 167–177. 2, 5

[LWGM04] LLOYD D. B., WENDT J., GOVINDARAJU N. K., MANOCHA D.: CC Shadow Volumes. In *EGSR* (2004), pp. 197–206. 1

[NRH03] NG R., RAMAMOORTHI R., HANRAHAN P.: All-frequency shadows using non-linear wavelet lighting approximation. *ACM Trans. Graph. 22*, 3 (2003), 376–381. 1

[Ots79] OTSU N.: A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern. 9*, 1 (1979), 62–66. 5

[PJH16] PHARR M., JAKOB W., HUMPHREYS G.: *Physically based rendering: From theory to implementation*, Third ed. Morgan Kaufmann, 2016. 1, 3

[RBA09] REINBOTHE C. K., BOUBEKEUR T., ALEXA M.: Hybrid ambient occlusion. In *Eurographics (Areas Papers)* (2009), pp. 51–57. 1

[RFB15] RONNEBERGER O., FISCHER P., BROX T.: U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI* (2015), pp. 234–241. 2

[SCHG23] SUN W., CHE Y., HUANG H., GUO Y.: Neural reconstruction of relightable human model from monocular video. In *ICCV* (2023), pp. 397–407. 1, 2, 5

[SLZ*22] SHENG Y., LIU Y., ZHANG J., YIN W., OZTIRELI A. C., ZHANG H., LIN Z., SHECHTMAN E., BENES B.: Controllable shadow generation using pixel height maps. In *ECCV* (2022), pp. 240–256. 2

[SSSJ20] SAITO S., SIMON T., SARAGIH J., JOO H.: PIFuHD: Multi-level pixel-aligned implicit function for high-resolution 3D human digitization. In *CVPR* (2020), pp. 84–93. 2, 3, 4, 6

[SWP11] SCHERZER D., WIMMER M., PURGATHOFER W.: A survey of real-time hard shadow mapping methods. *Comput. Graph. Forum 30*, 1 (2011), 169–186. 1

[SZP*23] SHENG Y., ZHANG J., PHILIP J., ET AL.: PixHt-Lab: Pixel height based light effect generation for image compositing. In *CVPR* (2023), pp. 16643–16653. 1

[TKE21] TAJIMA D., KANAMORI Y., ENDO Y.: Relighting humans in the wild: Monocular full-body human relighting with domain adaptation. *Comput. Graph. Forum 40*, 7 (2021), 205–216. 2, 5

[TTM*22] TEWARI A., THIES J., MILDENHALL B., ET AL.: Advances in neural rendering. *Comput. Graph. Forum 41*, 2 (2022), 703–735. 2

[VSP*17] VASWANI A., SHAZEER N., PARMAR N., USZKOREIT J., JONES L., GOMEZ A. N., KAISER Ł., POLOSUKHIN I.: Attention is all you need. pp. 5998–6008. 4

[ZLW19] ZHANG S., LIANG R., WANG M.: ShadowGAN: Shadow synthesis for virtual objects with conditional adversarial networks. *Computational Visual Media 5*, 1 (2019), 105–115. 4

[ZLWY23] ZHENG R., LI P., WANG H., YU T.: Learning visibility field for detailed 3D human reconstruction and relighting. In *CVPR* (2023), pp. 216–226. 1, 2, 5

Appendices

Appendix A: Low order Spherical Harmonics Reconstruction

Figure 8 shows the low order SH reconstruction of a directional illumination signal (represented in latitude-longitude 2:1 illumination map) cannot precisely localise the light source for two different solid angles. The original light source edges are blurred out and reconstruction artefacts are present with lower intensity in both reconstructions.

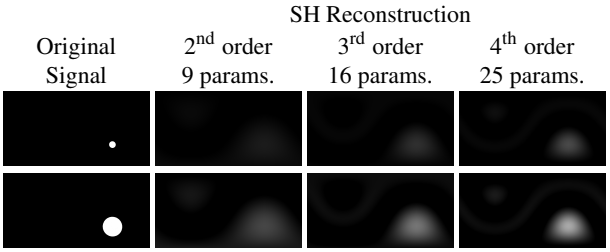

Figure 8: *Low order SH reconstruction of a directional illumination signal. Reconstructions in the first row are magnified by factor of 3 for visualisation purposes.*