Eurographics Symposium on Rendering (2023)
T. Ritschel and A. Weidlich (Editors)

Learning Projective Shadow Textures for Neural Rendering of Human Cast Shadows from Silhouettes

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 contribution introduces a two-step, novel neural rendering framework to learn the transformation from a 2D human silhouette mask to the corresponding cast shadows on background scene geometries. In the first step, the proposed neural renderer learns a binary shadow texture (canonical shadow) from the 2D foreground subject, for each point light source, independent of the background scene geometry. Next, the generated binary shadows are texture-mapped to transparent virtual shadow map planes which are seamlessly used in a traditional rendering pipeline to project hard or soft shadows for arbitrary scenes and light sources of different sizes. The neural renderer is trained with shadow images rendered from a fast, scalable, synthetic data generation framework. We introduce the 3D Virtual Human Shadow (3DVHshadow) dataset as a public benchmark for training and evaluation of human shadow generation. Evaluation on the 3DVHshadow test set and real 2D silhouette images of people demonstrates the proposed framework achieves comparable performance to traditional geometry-based renderers without any requirement for knowledge or computationally intensive, explicit estimation of the 3D human shape. We also show the benefit of learning intermediate canonical shadow textures, compared to learning to generate shadows directly in camera image space. Further experiments are provided to evaluate the effect of having multiple light sources in the scene, model performance with regard to the relative camera-light 2D angular distance, potential aliasing artefacts related to output image resolution, and effect of light sources' dimensions on shadow softness.

CCS Concepts

• ***Computing methodologies*** → ***Computer graphics; Neural networks;***

1. Introduction

Compositing techniques traditionally require, among other information, 3D geometry of foreground objects to render plausible shadows into the background plates of a *target* scene [Wri13]. In this context, *harmonising* the appearance (shading and/or shadows) of foreground objects without knowledge of geometry, or other scene properties such as materials, lighting, or camera is desirable and has recently received increased attention [SLZ*22, ZLZ*20, LLZ*20, SZB21, HNZ22].

A specific category of foreground objects with particular importance are human performers or presenters, extensively used with compositing techniques in the VFX industry. However, reconstructing the non-rigid human bodies in real-time or tracking a respective pre-calculated 3D model can be computationally expensive, or might simultaneously require multiple calibrated capturing devices.

In this paper, we specifically address the problem of hard shadow generation for *standing* human postures cast on *arbitrary* scene geometries, given the segmentation mask of the body from a single viewpoint of a *calibrated* camera. The proposed approach therefore does not require the 3D geometrical model of the subject and is not limited to certain shadowed scene geometry, e.g. a floor. However, as both the camera pose, and light source position are given with respect to the subject, our approach stands between the appearance harmonisation techniques [ZLZ*20, LLZ*20] and traditional rendering. Refer to Fig. 1 for an overview of our method.

A main assumption in the state-of-the-art neural appearance harmonisation is an existing *planar* scene geometry for shadows e.g. a floor in the background image (backplate) – either explicit [SZB21, LLZ*20], or implicit, by the nature of employed training/test datasets [ZLZ*20, HNZ22]. Although, given enough information about the scene, planar shadows generated in camera image space can be re-cast to a different background geometry, we propose and justify a two-step approach to shadow casting. In the first step, a *binary* shadow texture is generated for the foreground by the proposed neural model, from the view point of light source, similar to rendered depth maps in shadow mapping [Wil78], which in the second step are projected/cast to scene geometry using conventional rendering pipelines. In essence, both the rendering of a depth map for the foreground object from light source perspective, and the consecutive depth test for background scene geometry

DOI: 10.2312/sr.20231125

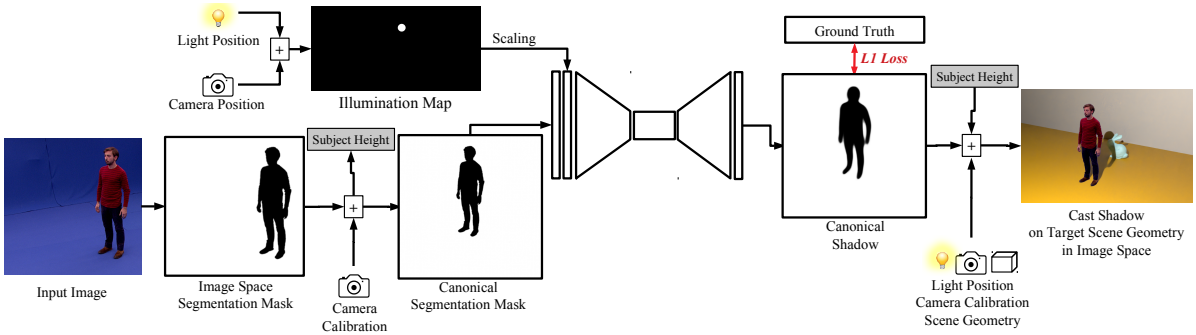


Figure 1: *Overview of our method for casting foreground subject's shadows on the background geometry: the inputs to our method are the normalised, canonical segmentation mask using the camera calibration (intrinsics), and a synthesised environment light based on the relative position (extrinsics) of camera and light to the subject. The output is the projective shadow texture (canonical shadow) which is cast on given, arbitrary background geometry using the estimated subject height, and the light position*

are learned in a binary projective shadow texture, which for convenience, we call *canonical shadow* hereafter.

Recently, Sheng et al. [SLZ*22] propose to learn an intermediate 2.5D *Pixel Height* representation from 2D foreground object mask, which later can be used with ray tracing to cast hard shadows in the image space. The generated hard shadows are then softened by a separately trained soft shadow generator (SSG [SLZ*22], SSG++ [SZP*23]) neural model. In contrast to soft shadow network (SSN) [SZB21], which learns the complex mapping from environment light maps and object silhouettes to soft shadows, SSG aims to soften the hard shadows by a softness factor. However, to support general shadow receivers of different geometries, one needs to have knowledge of, and integrate Pixel Height maps of the shadow receiving geometry in a traditional pipeline, which limits the applicability of the approach. Our proposed canonical shadows can seamlessly be used in a traditional rasterisation, or global illumination rendering pipeline, using the conventional representations of the background geometry such as triangular meshes.

The most important information for estimating cast shadows of an object is the 2D occluding contour from the light source perspective [Wil78]. In this paper, without 3D geometry and given only the subject's 2D occluding contour from the viewpoint of a calibrated camera, we learn its canonical shadow for an arbitrary light source location. Here the effects of 3D human shape on the canonical shadow are learnt implicitly from the training data. To render canonical shadows, in ray tracing algorithms, the binary textures are mapped to virtual, transparent (canonical) planes which are defined by the normal vector light source-subject position. These occluders are added to the scene to naturally generate hard or soft shadow effects on the background scene geometry. In rasterisation pipelines, the canonical shadows are mapped to existing scene geometry via projective texture mapping [SKVW*92] from the light source viewpoint. Note that separately generated shadows caused by individual light sources can be combined to model complex scene illumination.

A deep convolutional neural network (CNN) [GBC16] model is introduced and trained on a novel synthetic dataset generated on 311 human models of a wide variety of body and clothing characteristics which will be released publicly as 3DVHshadow. We show that the trained model has the generalisation capability to regress the shape of cast shadows on a test set of 107 models, and human silhouettes from real images. This generalisation is due to the geometrical aspects of the light transport that can be reproduced by synthetic data, in contrast to the learning and reproduction of the global illumination effects for relighting problems.

We further investigate how the proposed canonical shadows improve the details of the generated shadows, quantitatively and qualitatively, compared to both (a) directly generating the shadows in the camera image space on a planar background scene geometry, and (b) the shadows generated from the deep learning-based monocular 3D human geometry estimation method PIFu [SHN*19]. We ablate our model with regard to resolution to show it can maintain quality for larger output images, and therefore avoid unwanted aliasing effects. Additionally, we analyse the effect of 2D angular distance between light source and the camera in the quality of the generated shadows. This experiment shows the benefit our approach in implicitly learning the 3D geometry of the subject, for more demanding scenarios where light source and camera are perpendicular.

Finally, we show the generalisation of our method to images of real people by comparing the generated shadows by the proposed model to the shadows rendered using classic multiple-view 3D geometry reconstruction. The experiments show the suitability of the proposed model in regressing the shape of the shadows from the 2D silhouette only, without having 3D human geometry.

Our main contributions are as follows:

- An intermediate, projective hard shadow texture representation (canonical shadow) with seamless integration in conventional rendering pipelines;
- A corresponding neural rendering model for generation of such representation from 2D silhouette masks without 3D shape;
- A novel dataset 3DVHshadow for training and evaluation of human shadow generation, including a wide variety of people, clothing and poses;
- Plausible shadow generation and compositing for images of real people using traditional rendering pipelines.

Table 1: *Overview of related work for neural rendering of shadows. 'I' denotes no explicit assumption, but existing shadows on planar geometry predominantly occurring in the dataset*

	Shading	Shadow	Object	Training Dataset	Lighting	Existing Shadows	Arbitary S. Geom.
Wang et al. [WWL19]	✓	✓	2 Objects	Synthetic	Point	×	×
ShadowGAN [ZLW19]	×	✓	22 Categories	Synthetic	Point	✓	×
Zhan et al. [ZLZ*20]	✓	✓	Car, Human	Real	Directional	I	I
ARShadowGAN [LLZ*20]	×	✓	13 Categories	Mixed	Point/Sun	✓	×
SSN [SZB21]	×	✓(Soft)	43 Humans, 59 O.	Synthetic	Env. Map	×	×
Hong et al. [HNZ22]	×	✓	General	Real	Directional	I	I
SSG (Hard Sh. Phase) [SLZ*22]	×	✓	Humans	Mixed	Point	×	✓
Ours	×	✓	Diverse Humans	Synthetic	Point	×	✓

2. Related Work

Various aspects of illumination estimation and relighting have been studied in the literature. For a comprehensive review of approaches using neural models, refer to the recent survey by Einabadi et al. [EGH21]. Also, for an overview of the recent advances on neural rendering refer to Tewari et al. [TTM*22]. In this section, we review recent work which is most closely related to the problem of neural rendering of shadows.

Explicit Lighting Estimation. In this scenario, the scene lighting is explicitly estimated to insert additional objects for which the geometry and the reflection models are known beforehand. In this case, the shading and shadows of the inserted objects are readily available and their quality is dependent on the accuracy of the estimated parameters of the respective lighting model. Deep neural models have shown promising results for estimating the lighting of indoor [GHS*19, SF19], outdoor [HGAL19, ZSHG*19] or general [LMF*19, CSC*18] scenes as well as specific object(s) [PPeYW20], human faces [CLG*18], hands [MCV18], etc. The same is also valid in neural inverse rendering for estimating the scene elements, including the lighting parameters [SGK*19, LSR*20, YME*20]. It is however noteworthy to mention that even with realistic lighting and reflection models of a scene, it remains a challenge to simulate global illumination effects, sub-surface scattering, etc. with the current state-of-the-art physically-based renderers in a reasonable amount of time [PJH16].

Foreground-Background Harmonisation. A more challenging problem formulation is to harmonise the shading and shadows of a foreground object to a target scene, without necessarily having the respective geometry and the material models. Table 1 provides an overview of the related contribution characteristics.

A common approach is to generate the shadow(s) directly in the background image with the assumption of existing planar scene geometry. In case of Zhan et al. [ZLZ*20] and Hong et al. [HNZ22], there is no explicit assumption in the methodology, but the training/test datasets predominantly contain cast shadows on planar surfaces. However, more often in compositing tasks, the backplate has complex geometry and the shadow might not cast on a planar surface. Our contribution utilises the camera calibration to regress the canonical shadows which can then be cast to arbitrary known background geometry.

Zhan et al. [ZLZ*20] and SSN [SZB21], similar to our method, provide control over target lighting whereas other work [WWL19, ZLW19, LLZ*20, HNZ22] are dependent on and therefore limited to the existing cues in the backplate – e.g. visible occluders and their respective shadows – to generate shadows for the inserted foreground objects. This can prove challenging; e.g. ARShadowGAN [LLZ*20] reports that the proposed model fails to generate correct shadows where there exist large dark areas in the backplate. Also, Hong et al. [HNZ22] report larger shadows when there are no existing shadow cues in the backplate.

Zhan et al. [ZLZ*20], SSN [SZB21] and Hong et al. [HNZ22] address shadow generation for non-rigid human postures but their datasets are not diverse. Other work focus on rigid-body object categories or even instances [WWL19]. Our work proposes 3DVHshadow dataset containing variety of people, clothing and postures.

In contrast to other contributions, SSN [SZB21] generates only soft shadows. In this work, we however address the problem of rendering the shadow shape detail corresponding to the human subject, giving a realistic hard shadow mask for point light sources. We show that, by using canonical shadows, soft shadows can still be rendered and cast on arbitrary non-planar geometry for the scenes where light sources have physical geometry. The canonical shadows therefore enable rendering both detailed, and soft shadows for typical scene illuminations.

Shadow Casting by Intermediate Representations. Shadow mapping [SWP11, Wil78] is a successful, intermediate depth (buffer) representation rendered from the light source perspective which is thoroughly studied in computer graphics literature for casting shadows in arbitrary scenes using a depth test. Inspired by this straightforward technique, our proposal aims to learn this depth test for a certain foreground subject as a binary texture (canonical shadow) on a canonical virtual plane.

Recently, Griffiths et al. [GRP22] (OutCast) proposed to learn a non-binary shadow image for source, and target lighting directions to relight *outdoor* scenes with cast shadows. The proposed representation is the binary shadow mask of the scene multiplied by the cosine term, i.e., the "clamped dot product of the light direction and the approximate normal image," the latter to support self shadowing. This contribution demonstrates that the proposed representation is learnable from a coarsely estimated depth map obtained from a single input colour image.

SSG by Sheng et al. [SLZ*22] proposes to estimate an intermediate 2.5D Pixel Height representation from 2D foreground subject masks in the first step, and ray trace the shadows in the image space in the second step given the point light position. Our approach follows a two-step manner, too, but our proposed, intermediate canonical representation is readily integrable in existing traditional pipeline in a sense that conventional mesh geometry of the background scene is used, as opposed to calculating, and integrating Pixel Height maps in the rendering process.

SSG++ [SZP*23] claims "soft shadow integration [of Sheng et al. [SLZ*22]] in classical rendering algorithms is slow", however, it aims to mitigate this issue by multiple, *pre-calculated* input buffers containing Pixel Height maps of background, and foreground from multiple samples on the light source to generate soft shadows and other lighting effects. Pre-calculating such maps is an additional computational load in real-time rendering applications [SLZ*22].

Monocular/Multi-view 3D Human Geometry Estimation. In principle, 3D mesh geometries of foreground subjects can be estimated from input foreground cut-out (not mask) using deep neural models [SHN*19, SSSJ20, CLZL22], to be used in a conventional rendering algorithm. Such approaches show generally lower quality estimation for feet contact points [SLZ*22]. Another main issue is their high computational complexity. Our approach is a few orders of magnitude faster to estimate canonical shadows compared to monocular 3D mesh estimators.

From classic, non-neural methods, the 3D geometry of human subjects can be estimated using multiple-view optimisation techniques [BHKH13, SH07]. However, the camera setup is costly and not applicable to real-world scenarios where only monocular views are available.

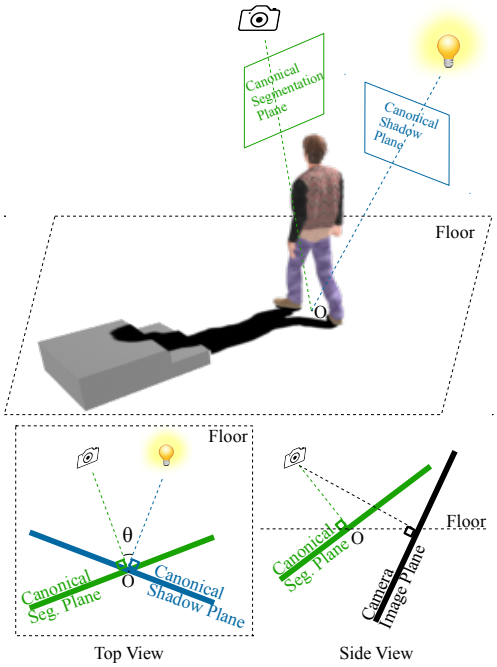


Figure 2: *A sample scene with the proposed canonical shadow and segmentation planes, respectively oriented towards the light source and the camera. The foreground subject is composited on the top layer after rendering the scene/shadow. The top view depicts the 2D angular distance between the camera and the light source, θ, with respect to subject position O. The normal of canonical segmentation plane is the vector connecting O to camera position as shown in the side view*

3. Neural Rendering of Shadows from 2D Human Silhouettes

The proposed human shadow generation model is a CNN trained on synthetic data with high geometric variety which is able to leverage the canonical shadow representation to maximise modelling quality. In this section, we first describe the proposed canonical shadow representation inspired by shadow mapping [Wil78], and the corresponding rendering pipeline. Then our neural model, including the training and implementation details are presented. Finally the synthetic data generation framework is presented in detail. Fig. 1 presents an overview of the approach.

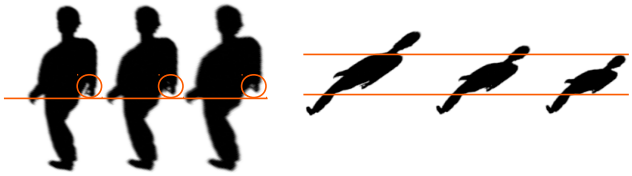
3.1. Canonical Shadow Representation: Projective Binary Hard Shadow Textures

In rendering, shadows of an object are generated by checking if the rays emitted from a light source to the scene geometry are occluded by the object. The proposed canonical shadow is a binary texture on a metric virtual (canonical) plane and represents this occlusion check for a certain object, given the relative light source position. The canonical shadow plane is defined partially by its normal, the vector from the light source to the subject position. The exact 3D position of the canonical plane on its normal is compensated for by the metric size of the canonical texture. In other words, the canonical shadow is effectively the predicted silhouette of the object as seen from the light source position projected on the canonical plane.

This representation is motivated from and simplifies the shadow mapping [Wil78] process for our specific shadow casting task; The binary texture of canonical shadow is analogous to the outcome of depth test against the depth map rendered from the light source perspective, assuming the scene has only two layers: the foreground (with no cast shadows on), and the background which receives the cast shadow from the former. Fig. 2 depicts the geometry of canonical planes for a sample scene.

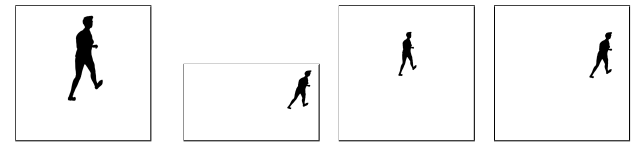
By choosing to learn canonical shadows, we ensure the existing benefits of traditional shadow mapping process in conjunction with our proposed neural rendering approach, compared to the shadows generated directly in the camera image space:

- It *facilitates* the learning of the shadow transformation process, i.e., the quality and details of the generated shadows are improved. This is due to the fact that shadow texture rendered from the light source perspective provide a compact proxy representation, independent of scene geometry, whereas for image space representation, i.e. on the ground plane, there are large variations in shadow scale and direction (Fig. 3).



(a) Canonical Space (b) Image Space

Figure 3: *(a) The metric canonical shadows for three different light source heights compared to (b) the corresponding shadows in image space on a ground plane. The changes in the former are minimal. Overlays provided for comparison*



Canonical Space Image Space

Figure 4: *Input normalisation: canonical format (left) for various camera image segmentation masks (right)*

- It substitutes the network's fixed-size image space output shadow in pixels with a metric shadow texture of estimated dimension. The canonical representation is independent of the scene configuration, background geometry, and the camera location, and, in contrast to image space representation, can always accommodate the whole shadow.

Similarly, we define the canonical segmentation plane with its normal, the vector connecting the camera to the subject position (Fig. 2). The benefit of such representation is when the segmentation mask is projected to the canonical plane and is size-normalised, its shape is agnostic to camera intrinsic parameters such as focal length, pixel size, etc., as well as the subject's position in the image. This transform is a planar homography where the camera position is the centre of projection as described below. Fig. 4 shows such input normalisation for a few segmentation masks.

3.2. Rendering Canonical Shadows

Rendering scenes with canonical shadows requires two phases: (a) rendering the background scene and the cast shadow(s), and (b) compositing the foreground subject cut-out given the provided segmentation mask. The latter is a straightforward compositing technique and is normally achieved by alpha blending operators. The first phase, however, can be performed either in rasterisation pipelines, or with ray tracing methods.

For rasterisation, rendering canonical shadows is equivalent to projective mapping of the 2D canonical shadow texture on the existing scene geometry from the corresponding light source viewpoint – first introduced by Segal et al. [SKVW*92] for rendering fast hard shadows using graphics hardware. To render real-time soft shadow effects, multiple techniques have been proposed in the literature [HLHS03]. For example, the simplest image-based method, introduced by Heckbert and Herf [HH97], is to render soft shadow

Table 2: *The proposed CNN architecture for shadow generation*

Module	Layer	Kernel	Resample	Output
	Conv	3×3	-	$512 \times 512 \times 4$
Encoder	DSL$\times$5	3×3	AvgPool 2×2	$32 \times 32 \times 64$
Shadow Trans.	ShL$\times$4	15×15	-	$32 \times 32 \times 64$
Decoder	USL$\times$4	3×3	Upsample $\times 2$	$512 \times 512 \times 8$
	Conv	1×1	-	$512 \times 512 \times 1$
	tanh	-	-	$512 \times 512 \times 1$

effects by sampling the geometry of light source at multiple positions to create binary occlusion maps (based on the given canonical shadow texture) and combine them in a soft shadow texture to be projected on the shadow receiving geometry.

On the other hand, for ray tracing algorithms, the 2D canonical shadow is mapped to a corresponding canonical plane and added to the scene as an occluder, to be rendered with existing background scene geometry. In photography or theatrical lighting, this shadow generation technique is usually referred to as placing *gobos* in front of a light source. For example, in path tracing [PJH16], both hard or soft shadows effects are naturally achieved based on the light source geometry, and other scene configurations. The only consideration is to make the occluder invisible to the camera rays to hide it in rendered images.

To render soft shadows cast from geometrical light sources, canonical shadows are generated corresponding to a representative point on the light source geometry, e.g., the centre for spherical lights. This is an acceptable approximation for the estimated shadow masks so long as the light source dimensions over distance are negligible compared to the subject size – otherwise, the light source must be sampled by multiple point light sources, corresponding to multiple occluder canonical shadows.

If there are multiple light sources present in the scene, rendering canonical shadows in phase (a), is achieved in a layered manner: Each rendering layer has one light source, and the respective canonical plane/shadow. The rest of the scene is shared between all layers. All layers are rendered separately, and then *added* together to form the final image. This is based on rules of superposition [NSD95], i.e., "the linearity of the rendering operator with respect to illumination for a fixed scene and camera geometry".

In our implementations, we use Blender's path tracing engine Cycles [Ble] for rendering single or multiple, hard or soft canonical shadows. Appendix A describes the rendering process in detail.

3.3. Shadow Generation Model

The encoder-decoder architecture of the proposed CNN model is inspired by the related work [HNZ22, ZLW19] and is presented in detail in Table 2. Each downsampling (DSL), upsampling (USL), or Shadow Transformation (ShL) layer consists of a convolution, ReLU non-linearity, and instance normalisation. The Shadow Transformation module in the latent space benefits from larger receptive field convolutions to take into account the global nature of the shadow generation transformation. The tanh in the last layer limits the output range of each pixel so that the shadow image can be obtained by a range normalisation. The model has 3.8 million parameters. Appendix B provides a detailed description of the in-

Table 3: *Quantitative metrics calculated on the synthetic generalisation dataset for the baselines and the proposed canonical method. The shadow output of the neural renderer is thresholded by the specified pixel percentage values for IoU and Dice metrics*

			$< 95\%$				$< 99\%$			
	RMSE ↓		IoU ↑		Dice ↑		IoU ↑		Dice ↑	
Method	avg	std	avg	std	avg	std	avg	std	avg	std
Baseline (a) Shadows by Planar Homography	0.159	0.057	0.617	0.208	0.739	0.192	0.625	0.209	0.745	0.192
Baseline (b) Image Space Shadows	0.153	0.032	0.683	0.096	0.807	0.074	0.696	0.093	0.817	0.070
Baseline (c) Shadows from PIFu Geometry	0.221	0.062	0.551	0.148	0.698	0.135	0.565	0.147	0.710	0.132
Canonical Shadows	**0.097**	0.028	**0.827**	0.059	**0.904**	0.037	**0.835**	0.056	**0.909**	0.035

Figure 5: *An example entry of 3DVHshadow: (a) dataset's diverse human models, (b) a textured human with (c) its segmentation mask, and (d) the corresponding shadow mask on a planar surface*

put channels (the segmentation mask, and illumination), as well as the implementation details.

Loss Function. The loss criteria used in the back-propagation training algorithm is an L1 reconstruction loss, calculated on the network's output pixel values, $\hat{I}$, and the corresponding supervision label, I, as $\mathcal{L}_{rec} = \sum_{trainingset} |\hat{I} - I|$.

3.4. Synthetic Data Generation Framework

To train and evaluate the proposed neural rendering model, a diverse, synthetic dataset is generated based on a set of 3D human models of 3DVH virtual human dataset [CMIH21] performing various walking styles such as while answering a call, happy walking, etc. This extension to 3DVH will be released as 3DVHshadow to facilitate future research on this topic.

Each dataset entry includes a rendering of the subject from the camera view point, its binary segmentation mask, and a binary shadow mask on the floor where the subject stands – in total 3 images (Fig. 5). The respective rendering metadata such as point light source position, camera pose, etc. is also provided alongside the images. In total, the dataset contains about 24400 training and 8400 test entries, each rendered with random postures, and camera and light source positions. Appendix C provides further details on the employed 3D models and rigging, scene content, rendering algorithm and its settings, and the factors of variations.

4. Experiments

In the following, we evaluate the generalisation of the trained model in the proposed canonical shadow space for unseen combinations of body shape, posture, camera pose and point light source positions, on synthetic and real images.

4.1. Synthetic Data

In this experiment, we use the introduced synthetic training dataset to train the proposed model. For evaluation, we use the corresponding synthetic generalisation set.

We also use the same dataset to train the same encoder-decoder model to generate shadows in the camera image space on a predefined geometry (here a floor) and examine the quality of the generated image space shadows versus the proposed canonical representation.

Baselines Methods. We consider three baseline methods for comparisons: shadows generated on a planar surface by (a) a naive planar homography transformation applied to segmentation masks, assuming that the subject is planar, i.e. has no depth; (b) our proposed model, but trained in the camera image space (not canonical) for segmentation, and shadow masks; (c) based on the estimated subject's 3D geometry from PIFu [SHN*19]. In all three cases the generated shadows are then converted into the canonical form for comparison. Evaluation in canonical shadow space is agnostic to scene geometry.

The baseline method (a) is based on the property that the image of a planar object and the image of its shadow are related through a homography. Camera calibration and light position are employed to estimated this projective transformation by obtaining the required 4-point correspondences on the planar shadow receiver for the corners of the 2D bounding box of the segmentation mask.

For baseline (c), for fairness, we trained the PIFu model based on our synthetic dataset using the official code [SHN*19]. Appendix D covers the PIFu's training parameters and test details, as well as the details of the input height normalisation step for baseline (b).

Metrics and Quantitative Results. We calculate Root Mean Square Error (RMSE), Intersection over Union (IoU), and Dice (F1 Score) [SM83] metrics for the generated shadows by each method, comparing them to the shadow labels of the test dataset which are based on ground truth 3D geometry. As the output of the neural renderer is real-valued pixels in the range of $[0,1]$, the IoU and Dice metrics are calculated on thresholded, binary shadows. Table 3 shows the results for the proposed method trained in the canonical space, compared to the aforementioned baselines. Our method outperforms the baselines for all RMSE, IoU, and Dice metrics and

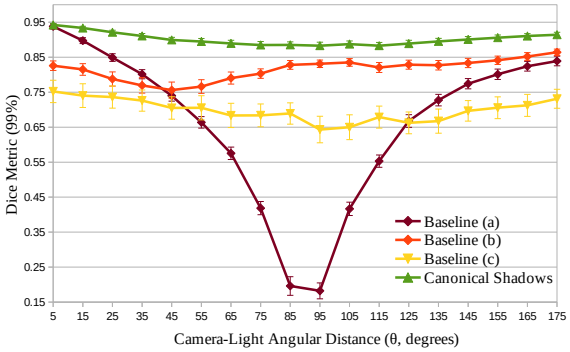


Figure 6: *Dice metric calculated on 18 10-degree bins of 2D camera-light angular distance (θ in Fig. 2). Standard deviation bars are linearly scaled (25%) for clear visualisation*

has the lowest standard deviation. The standard deviation is consistently the highest for baseline (a) for all metrics.

Fig. 6 depicts the Dice metric calculated on 18 bins (10 degrees each) of 2D camera-light angular distance which is referred to as θ in Fig. 2. When the angle θ is near 0, all methods perform almost their best due to the fact that the expected shadow is very similar to the input segmentation mask, and therefore there exists little to none 3D extrapolation. Baseline (a) fails when θ is about 90 degrees (rows 2 and 6 of Fig. 7). This is due to the simplification that the shadows are generated for the planar (flat) subjects. Also, almost correct shadows are expected when angular distance is negligible (rows 3 and 7 of Fig. 7). This leads to higher standard deviation for baseline (a) for the metrics in Table 3. Our method performs mostly consistently when θ changes from 0 to 180 degrees.

Qualitative Results. Fig. 7 shows the results of the experiment for a number of samples randomly drawn from the generalisation set. The shadows from baseline methods are transferred to the canonical space to assist visual comparison. The results show more shape details for the generated canonical shadows compared to the corresponding baseline methods in image space. Baseline (c) demonstrates artefacts resulting from depth ambiguities in the monocular 3D mesh inference process in row 2, as opposed to rows 3 and 7 where its extrapolation performance is not challenged/visible due to specific camera and light perspectives. Also, baseline (c) is the only approach with detached shadow artefacts from the body. Baseline (b) suffers from lack of details, e.g. the missing limbs, in the generated shadows. However, the subject's height and overall shape are estimated.

Ablation to Output Resolution. Similar to shadow mapping techniques, projective shadow casting is prone to aliasing issues. One specific case is caused by upsampling smaller shadow textures for larger target rendering sizes. We ablate our proposed model with the same dataset but with higher resolution segmentation and shadow masks. Table 4 shows the metrics calculated for higher resolutions. The results demonstrate that increasing the convolution layers' kernel sizes according to the input/output resolution improves the shadow generation quality for higher resolutions compared to using the original kernel sizes of 3. This ablation is of particular importance while the input/output resolution can vary considerably in compositing tasks.

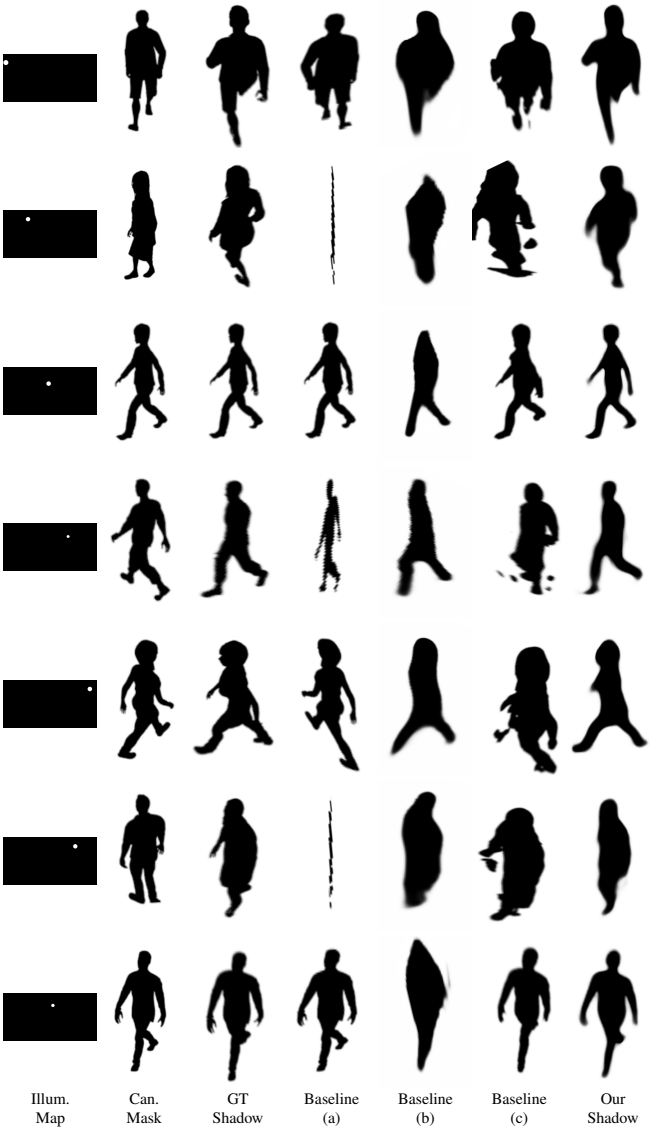


Figure 7: *Generated shadows for representative samples from the synthetic dataset. All shadows and the segmentation mask are shown on the canonical planes to assist comparison. Ground truth (GT) and baseline (c) shadows are rendered using traditional graphics pipeline and then transferred to canonical planes*

Moving Light Source. Additional material contains videos of subjects lit with moving (rotating) light source(s) with their shadows rendered using the generated canonical shadows compared to the ground truth. A reference object (Stanford Bunny [Sta]) is added for comparison purposes. The results demonstrate our approach does not suffer from artefacts related to temporal changes in the light source position.

Other Remarks. The same canonical shadow representation, given

Table 4: *Quantitative metrics calculated for learning canonical shadows using the same architecture as Table 2 but with higher input/output resolutions. Convolution layers' kernels are enlarged proportionally to the resolution*

				< 95%				< 99%			
		RMSE ↓		IoU ↑		Dice ↑		IoU ↑		Dice ↑	
Resolution	Kernel Size	avg	std	avg	std	avg	std	avg	std	avg	std
512 × 512	3	0.097	0.028	0.827	0.059	0.904	0.037	0.835	0.056	0.909	0.035
1024 × 1024	7	0.109	0.029	0.816	0.063	0.897	0.039	0.822	0.061	0.901	0.038
1024 × 1024	3	0.112	0.030	0.809	0.063	0.893	0.040	0.815	0.062	0.897	0.039
2048 × 2048	15	0.123	0.032	0.786	0.070	0.879	0.045	0.792	0.069	0.882	0.044
2048 × 2048	3	0.126	0.033	0.780	0.073	0.875	0.047	0.785	0.072	0.877	0.046

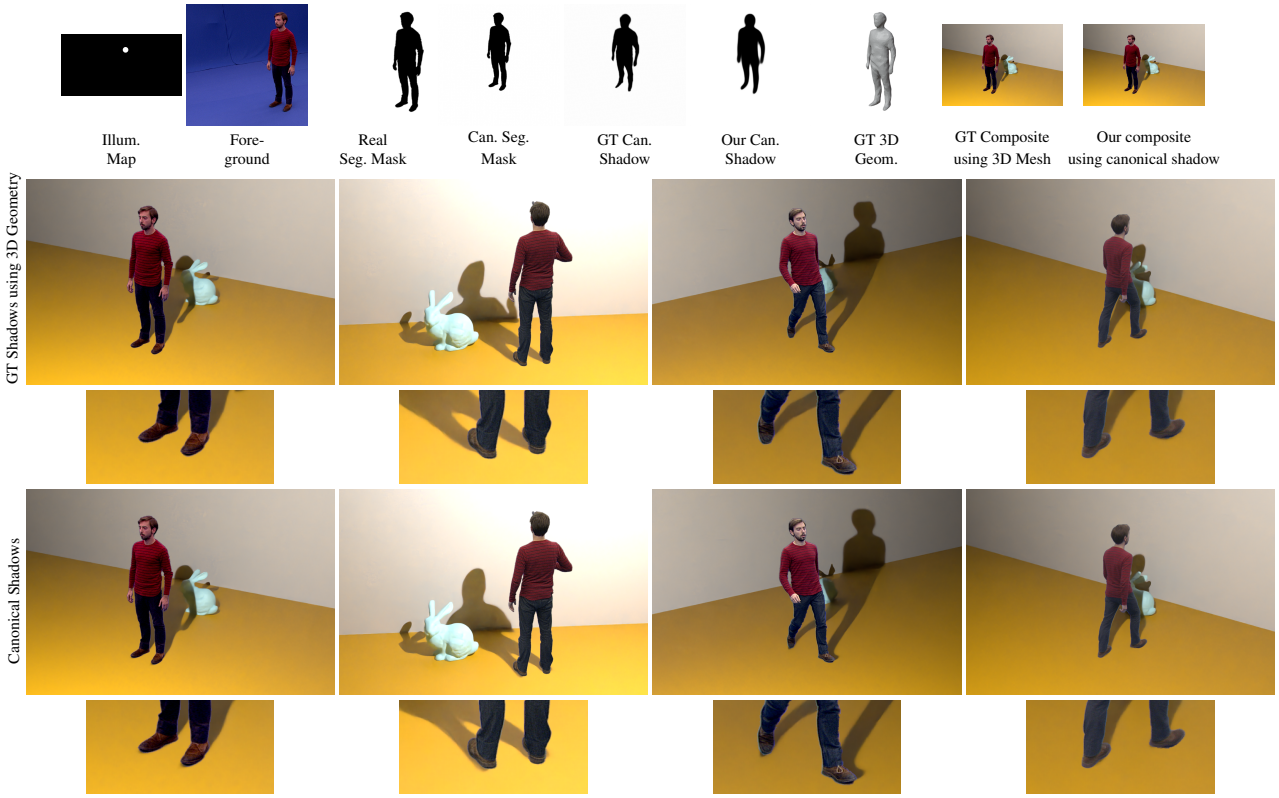


Figure 8: *Compositing canonical shadows of real silhouettes on the background scene (rows 4, 5) compared to the ground truth shadows from estimated 3D geometry (row 2, 3). Feet contact point shadows are zoomed in for better comparison. First row shows the pipeline data. Experiment performed on the frames of the Dan dataset [CVS]*

training data, can be used to learn shadow generation for other object categories, as the assumptions are not dependent on the human category.

The fast inference of canonical shadows (less than 5 ms for a mini-batch of size 8 for the resolutions in Table 4) makes the proposed approach suitable for compositing tools e.g. when a trained artist needs to interactively approximate the target scene lighting by trial and error based on the generated shadows. This is in contrast to Baseline (c) where estimating geometry from subject's cut-out requires about 5 seconds which is in principle 3 orders of magnitude slower.

It is also noteworthy that our main contribution is learning projective shadow textures (canonical shadows) for foreground subjects and comparing its benefits to the shadow generation in image space, or based on a monocular 3D geometry estimation method. These benefits include better quality and details, and the guaranteed whole shadows for the challenging non-rigid human object category (see Section 3.1). However, considering the general problem of neural rendering of shadows in the camera backplates, one in principle could compare the results to the related work, by training their models with our dataset, should they have the same inputs and outputs in the problem formulation. Based on Table 1, methods of Wang et al. [WWL19] and Zhan et al. [ZLZ*20] require the shading information of the backplate for the training and

generating shadows, while our formulation of the problem is mask-based and uses controllable lighting. ShadowGAN [ZLW19], AR-ShadowGAN [LLZ*20], and method of Hong et al. [HNZ22] require existing shadow-casting objects in the backplate. Hong et al. [HNZ22] use backplate shading information when there are no such existing objects. SSN [SZB21] and SSG [SLZ*22] are, similar to our method, mask-based and benefits from controllable lighting, but SSN by nature generates only soft shadows which are not comparable to our method's. SSG renders hard shadows in the first phase from the estimated Pixel Height maps, but the implementation details are not shared, and so the results are not reproducible at this stage.

4.2. Real Data

We compare the quality of the shadows generated by our neural model (second row, Table 4) to those generated by having the 3D mesh model of an actor in a studio, obtained from a baseline multiple-view 3D geometry reconstruction algorithm. More specifically, the selected *Dan*, *Character1*, and *J.P.* datasets [CVS] consist of multiple-view recordings of some basic human actions such as walking, moving a box, etc. from 8 calibrated cameras. The dataset entries are accompanied by the respective subject's position, segmentation mask, and the reconstructed geometry with the algorithms of Starck and Hilton [SH07], Budd et al. [BHKH13].

In this experiment, we render a synthetic scene from the point of view of the calibrated real camera and use the multi-layer rendering technique described in Section 3.2 to render canonical shadows and composite the foreground subject. The synthetic scene contains the subject's canonical shadow, an additional test object (Stanford Bunny [Sta]), floor, wall, and a randomly positioned point light source. The scene is rendered with Blender's Cycles engine [Ble] using the default settings.

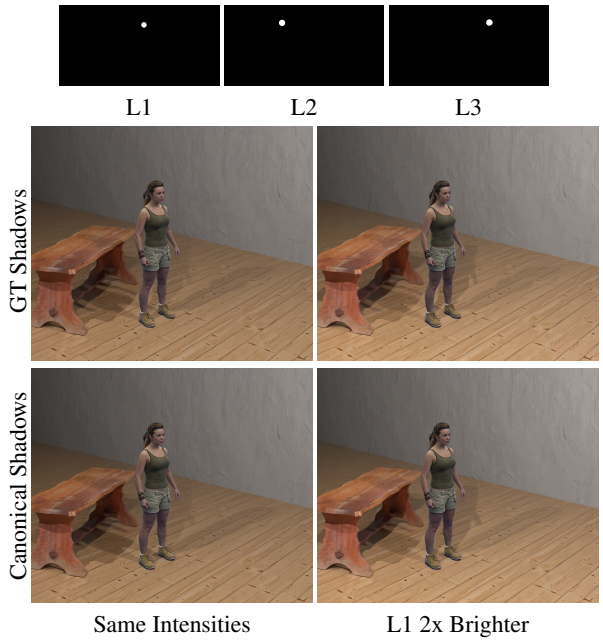


Figure 9: *Shadows in a scene with multiple point light sources. Ground truth shadows are generated using estimated 3D geometry. Experiment performed on the frames of Character1 from [CVS]*

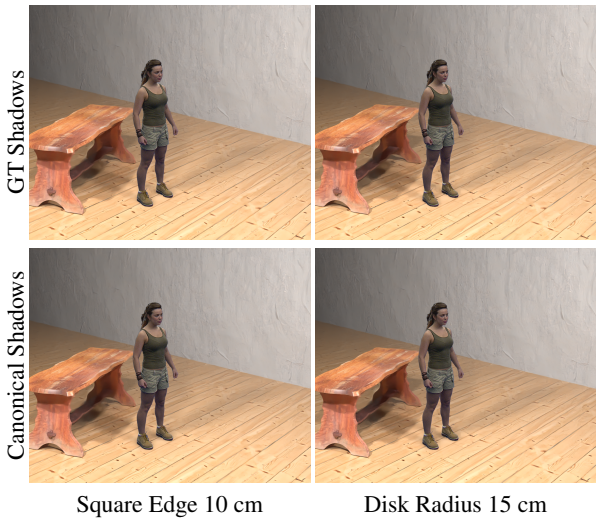


Figure 10: *Soft shadows generated for the scene in Fig. 9 given the geometry of physical light sources*

Fig. 8 shows the shadows generated by our proposed neural rendering approach. The direction, shape details and the scale of the shadows are plausible compared to the shadows rendered based on 3D geometry. Furthermore, this experiment shows the flexibility of the canonical shadows to be projected on arbitrary background scene geometries.

Scenes with Multiple Light Sources. Although, the proposed neural model assumes a single point light source as input, shadows from complex illuminations can be produced, e.g., by sampling multiple, discrete point light sources [Deb08] and generating and compositing the shadows. Note that this is different from sampling the emitting surfaces of light sources for rendering soft shadows [HLHS03].

Using our model, a canonical shadow is generated for each point light source separately. These canonical shadows are then rendered in a multi-layer 3D scene and composited using a conventional graphics pipeline (refer to Section 3.2) to obtain a scene with multiple shadows. Fig. 9 shows a scene rendered with three point light sources with various intensities.

Soft Shadows and Non-point Light Sources. Fig. 10 depicts the cast soft shadows for physical light sources with geometry, given the canonical shadows corresponding to the centre of the light sources. The rendering pipeline is discussed in Section 3.2, with the only difference that the scene light sources are modelled as emitting geometries rather than points. The results show that our approach renders visually plausible soft shadows on arbitrary scene geometry based on the estimated canonical shadows. Ground truth shadows using subject's 3D geometry are provided for visual comparison.

Note that the quality of the soft shadows are dependent on the quality of the estimated canonical shadows which is investigated

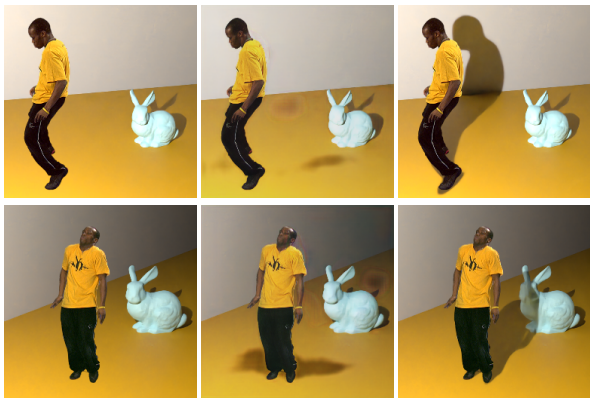


Figure 11: *ARShadowGAN [LLZ*20] shadows generated based on the naive composite vs. our canonical shadows. Experiment performed on the frames of J.P. from [CVS]*

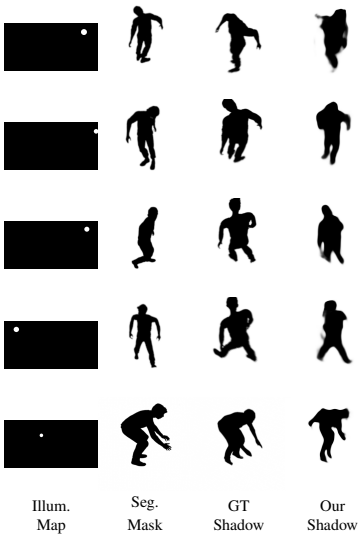


Figure 12: *Cases with the highest error value in the synthetic generalisation set (rows one to four) and a failure case on the real set (row five)*

in Section 4.1. The rendering algorithm and its settings can be changed for empirical reasons, e.g., time and quality constraints, etc. We use the path tracing algorithm of the Cycles engine [Ble] with the default settings according to the process described in Appendix A. This is in contrast to SSN [SZB21], which learns the soft shadow generation on only planar surfaces in one step.

Comparison with ARShadowGAN [LLZ*20]. ARShadowGAN [LLZ*20] uses existing shadow cues in the backplate to generate shadows for the inserted subject. Despite the fact that our proposed approach is based on controllable lighting, we aim to compare the methods by using the shadow cues of a background object – here the Stanford Bunny [Sta].

Fig. 11 shows our approach is capable of rendering detailed cast shadows on the background scene geometry and generates plausible shadows for humans. Comparisons demonstrate ARShadowGAN [LLZ*20] renders artefacts around subject's face and arms areas and on the scene geometry. Also, shadows are not plausible. The latter could fairly be associated with the lack of diverse human postures in the training dataset.

Compared to ARShadowGAN [LLZ*20], our approach uses extra camera calibration, subject's pose, background scene geometry, and light source position in a traditional rendering pipeline to (a) cast shadows on non-planar geometries, and (b) cast multiple shadows for multiple light sources. Liu et al. [LLZ*20] claim that "extending the Shadow-AR dataset is a possible way to solve [these] limitations." Also, our approach does not impose limitations on cast shadows intersecting the shadows of other existing geometries.

4.3. Limitations and Failure Cases

Fig. 12 shows the entries with the highest test error in the synthetic generalisation set as well as a failure case in the real dataset. For the synthetic ones, the worst cases' shadows are still plausible but they lack details.

The shadow generation model does not generalise to new human postures which are distinct from the poses seen in the training dataset such as the bending posture in Fig. 12 (bottom row). To overcome this limitation, the training dataset could be extended to include a wider range of human postures covering the full pose range.

5. Conclusion

In this paper, we proposed a two-step method using canonical shadow representation for neural rendering of human shadows on arbitrary background scene geometries, given the subject and light positions, subject's 2D silhouette mask, camera calibration, and without any knowledge of subject's 3D shape. The proposed model is trained with the synthetically generated dataset 3DVHshadow. The introduced dataset contains images, segmentation masks and shadows of subjects of various body characteristics and postures, lit and rendered under randomly sampled point light source and camera positions. The experiments show the benefits of training the neural renderer in the canonical, projective space, compared to directly in camera images, for generating higher quality cast shadows for synthetic, random scenes. The method is furthermore evaluated for generation of shadows for unseen images of real people and demonstrated to produce plausible shadows.

Future research directions include extending the range of human postures through extension of the training set, evaluating the role of different (visually inspired) loss functions, further analysis of aliasing through learning canonical shadows in normalized device coordinate space (perspective space), and evaluating temporal coherence of the generated shadow masks of moving subjects in video. Currently the proposed approach generates shadows for each image independently, which may result in flicker for video sequences. One approach is to explore contrastive loss with siamese networks to enforce temporal coherency in video, i.e., similar shadows for

neighbouring frames. This has the potential for a practical tool for artist driven shadow generation in media production.

Acknowledgements

We would like to thank Marco Pesavento for reproducing PIFu results on our dataset. This research was supported by UKRI EPSRC Platform Grant EP/P022529/1, and EPSRC BBC Prosperity Partnership AI4ME: Future Personalised Object-Based Media Experiences Delivered at Scale Anywhere EP/V038087/1.

Availability of data

The authors confirm that the datasets generated as part of this research are freely available under the terms and conditions detailed in the licence agreement enclosed in the data repository. Details of the data and how to obtain access are available from the University of Surrey: doi.org/10.15126/surreydata.900731 and the data repository website: cvssp.org/data/crh.

References

[Adoa] ADOBE: Fuse. adobe.com/products/fuse.html. 13

[Adob] ADOBE: Mixamo. mixamo.com. 13

[BHKH13] BUDD C., HUANG P., KLAUDINY M., HILTON A.: Global non-rigid alignment of surface sequences. *Int. J. Comput. Vis. 102*, 1-3 (2013), 256–270. 4, 9

[Ble] Blender project. blender.org. 5, 9, 10, 12, 13

[CLG*18] CALIAN D. A., LALONDE J.-F., GOTARDO P., SIMON T., MATTHEWS I., MITCHELL K.: From faces to outdoor light probes. *Comput. Graph. Forum 37*, 2 (2018), 51–61. 3

[CLZL22] CHAN K., LIN G., ZHAO H., LIN W.: S-PIFu: Integrating parametric human models with pifu for single-view clothed human reconstruction. *NeurIPS 35* (2022), 17373–17385. 4

[CMIH21] CALISKAN A., MUSTAFA A., IMRE E., HILTON A.: Multi-view consistency loss for improved single-image 3D reconstruction of clothed people. In *ACCV* (2021), pp. 71–88. 6, 13

[CSC*18] CHENG D., SHI J., CHEN Y., DENG X., ZHANG X.: Learning scene illumination by pairwise photos from rear and front mobile cameras. *Comput. Graph. Forum 37*, 7 (2018), 213–221. 3

[CVS] CVSSP3D DATASETS: Dan, Character1, and J.P. datasets. cvssp.org/data/cvssp3d. 8, 9, 10

[Deb08] DEBEVEC P.: A median cut algorithm for light probe sampling. In *SIGGRAPH Classes*. 2008, pp. 1–3. 9

[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. 3

[GBC16] GOODFELLOW I., BENGIO Y., COURVILLE A.: *Deep Learning*. MIT Press, 2016. 2

[GHS*19] GARDNER M.-A., HOLD-GEOFFROY Y., SUNKAVALLI K., GAGNÉ C., LALONDE J.-F.: Deep parametric indoor lighting estimation. In *ICCV* (2019), pp. 7174–7182. 3

[GRP22] GRIFFITHS D., RITSCHEL T., PHILIP J.: OutCast: Outdoor single-image relighting with cast shadows. *Comput. Graph. Forum 41*, 2 (2022), 179–193. 3

[HDL97] HORAUD R., DORNAIKA F., LAMIROY B.: Object pose: The link between weak perspective, paraperspective, and full perspective. *Int. J. Comput. Vis. 22*, 2 (1997), 173–189. 13

[HGAL19] HOLD-GEOFFROY Y., ATHAWALE A., LALONDE J.-F.: Deep sky modeling for single image outdoor lighting estimation. In *CVPR* (2019). 3

[HH97] HECKBERT P. S., HERF M.: *Simulating soft shadows with graphics hardware*. Tech. Rep. CMU-CS-97-104, Carnegie Mellon University, 1997. 5

[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. 5, 9

[HNZ22] HONG Y., NIU L., ZHANG J.: Shadow generation for composite image in real-world scenes. In *AAAI* (2022), p. 1360. 1, 3, 5, 9

[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). 1, 3, 9, 10

[LMF*19] LEGENDRE C., MA W.-C., FYFFE G., FLYNN J., CHARBONNEL L., BUSCH J., DEBEVEC P.: DeepLight: Learning illumination for unconstrained mobile mixed reality. In *SIGGRAPH Talks* (2019). 3

[LSR*20] LI Z., SHAFIEI M., RAMAMOORTHI R., SUNKAVALLI K., CHANDRAKER M.: Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image. In *CVPR* (2020), pp. 2472–2481. 3

[MCV18] MARQUES B. A. D., CLUA E. W. G., VASCONCELOS C. N.: Deep spherical harmonics light probe estimator for mixed reality games. *Comput. Graph. 76* (2018), 96 – 106. 3

[NSD95] NIMEROFF J. S., SIMONCELLI E., DORSEY J.: Efficient re-rendering of naturally illuminated environments. In *Photorealistic Rendering Techniques* (1995), pp. 373–388. 5

[NYD16] NEWELL A., YANG K., DENG J.: Stacked hourglass networks for human pose estimation. In *ECCV* (2016), pp. 483–499. 13

[PJH16] PHARR M., JAKOB W., HUMPHREYS G.: *Physically based rendering: From theory to implementation*. Morgan Kaufmann, 2016. 3, 5

[PPeYW20] PARK J., PARK H., EUI YOON S., WOO W.: Physically-inspired deep light estimation from a homogeneous-material object for mixed reality lighting. *IEEE Trans. Vis. Comput. Graph. 26*, 5 (2020), 2002–2011. 3

[SF19] SONG S., FUNKHOUSER T.: Neural illumination: Lighting prediction for indoor environments. In *CVPR* (2019), pp. 6911–6919. 3

[SGK*19] SENGUPTA S., GU J., KIM K., LIU G., JACOBS D. W., KAUTZ J.: Neural inverse rendering of an indoor scene from a single image. In *ICCV* (2019), pp. 8597–8606. 3

[SH07] STARCK J., HILTON A.: Surface capture for performance-based animation. *IEEE Comput. Graph. Appl. 27*, 3 (2007), 21–31. 4, 9

[SHN*19] SAITO S., HUANG Z., NATSUME R., MORISHIMA S., KANAZAWA A., LI H.: PIFu: Pixel-aligned implicit function for high-resolution clothed human digitization. In *ICCV* (2019), pp. 2304–2314. 2, 4, 6

[SKVW*92] SEGAL M., KOROBKIN C., VAN WIDENFELT R., FORAN J., HAEBERLI P.: Fast shadows and lighting effects using texture mapping. In *SIGGRAPH* (1992), pp. 249–252. 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. 1, 2, 3, 4, 9

[SM83] SALTON G., MCGILL M.: *Introduction to modern information retrieval*. McGraw Hill, 1983. 6

[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. 4

[Sta] Stanford Bunny. graphics.stanford.edu. 7, 9, 10, 12

[SWP11] SCHERZER D., WIMMER M., PURGATHOFER W.: A survey of real-time hard shadow mapping methods. *Comput. Graph. Forum 30*, 1 (2011), 169–186. 3

[SZB21] SHENG Y., ZHANG J., BENES B.: SSN: Soft shadow network for image compositing. In *CVPR* (2021), pp. 4380–4390. 1, 2, 3, 9, 10

[SZP*23] SHENG Y., ZHANG J., PHILIP J., HOLD-GEOFFROY Y., SUN X., ZHANG H., LING L., BENES B.: PixHt-Lab: Pixel height based light effect generation for image compositing. *arXiv preprint arXiv:2303.00137* (2023). 2, 4

[TTM*22] TEWARI A., THIES J., MILDENHALL B., SRINIVASAN P., TRETSCHK E., YIFAN W., LASSNER C., SITZMANN V., MARTIN-BRUALLA R., LOMBARDI S., ET AL.: Advances in neural rendering. *Comput. Graph. Forum 41*, 2 (2022), 703–735. 3

[Wil78] WILLIAMS L.: Casting curved shadows on curved surfaces. In *SIGGRAPH* (1978), pp. 270–274. 1, 2, 3, 4

[Wri13] WRIGHT S.: *Digital compositing for film and video*. Taylor & Francis, 2013. 1

[WWL19] WANG X., WANG K., LIAN S.: Deep consistent illumination in augmented reality. In *ISMAR Adjunct* (2019), pp. 189–194. 3, 8

[YME*20] YU Y., MEKA A., ELGHARIB M., SEIDEL H.-P., THEOBALT C., SMITH W. A. P.: Self-supervised outdoor scene relighting. In *ECCV* (2020). 3

[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. 3, 5, 9

[ZLZ*20] ZHAN F., LU S., ZHANG C., MA F., XIE X.: Adversarial image composition with auxiliary illumination. In *ACCV* (2020). 1, 3, 8

[ZSHG*19] ZHANG J., SUNKAVALLI K., HOLD-GEOFFROY Y., HADAP S., EISENMAN J., LALONDE J.-F.: All-weather deep outdoor lighting estimation. In *CVPR* (2019). 3

Appendices

Appendix A: Rendering Canonical Shadows with Cycles

Fig. 13 depicts the process of rendering a subject's shadow, given the canonical shadow and the corresponding point light source position, on a scene geometry, using Blender's path tracing engine Cycles [Ble].

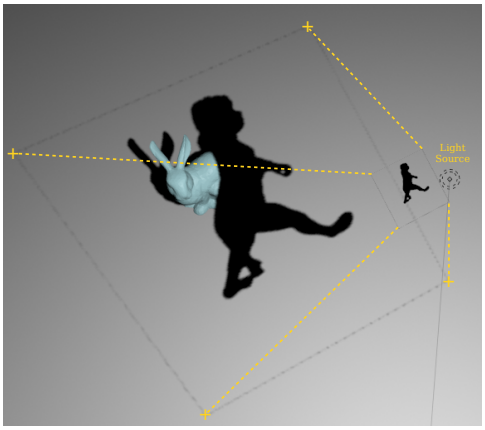


Figure 13: *Projecting a canonical shadow on the scene geometry. The yellow marks are the intersection points corresponding to the light rays passing through the canonical shadow bounding box corners. Note how the shadow is cast on the Stanford Bunny [Sta]*

In Cycles, this process is natively supported by assigning a transparent material to the canonical shadow plane (Fig. 13, smaller grey square) and having the canonical shadow as its texture. The canonical plane is set not visible to camera rays (hidden), but visible to the corresponding light source. This occluder is added to the existing scene and rendered with the rest of the scene elements.

Fig. 14 shows the rendering and compositing pipeline for multiple canonical shadows in Blender, based on the multi-layer rendering technique described in Section 3.2. In the last step, the subject is composited over the rendered image using the segmentation mask. This is a well established rendering process and can be automated, e.g., via Blender's Python scripting.

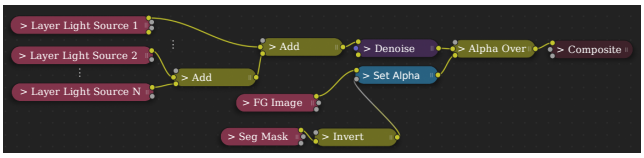


Figure 14: *Rendering and compositing pipeline for multiple canonical shadows in Blender*

Appendix B: Input Channels and Implementation Details

Table 2 presents the proposed model architecture. The network has two input channels: the binary, canonical segmentation mask of the subject, and the desired illumination under which the canonical shadow mask (the output) is generated (Fig. 1).

The Illumination Channel. It is modelled as a full spherical binary map scaled to have the same aspect ratio as the segmentation mask. The point light source is rendered as a homogeneous white blob on the map with a radius inversely proportional to its distance from the subject, multiplied by a size constant. This spherical map is then rotated around its axis according to the planar angular distance between the light source and the camera. Fig. 2 (top view) shows the required rotation as θ. Both segmentation and shadow masks are represented with black colour on a white background.

The Segmentation Mask Channel. Camera calibration is used to project the segmentation mask to the corresponding canonical plane, using the planar projective transformation described in Section 3.2. The subject on this metric plane is then cropped, and mapped to the network's input channel size (in pixels), using the estimated subject height. The generated output canonical shadow is then scaled back by the same height (see Fig. 1). We pad the masks proportional to the subject height to allow the canonical representation have full limbs in input and shadows.

More specifically, the inputs to the pre-processing are: foreground segmentation mask, and camera calibration i.e. intrinsic parameters, and pose w.r.t. the subject's contact point on the floor. The outputs are: subject's height, and canonical segmentation mask.

Note that if camera pose is not known w.r.t. the subject, but some other origin, given e.g. the camera height from the floor, the subject feet-floor intersection point can be easily calculated. This is achieved by intersecting the corresponding camera ray with the floor plane.

Implementation Details. The model is implemented in Python using the PyTorch library. The Adam optimiser is employed in the

training phase with the learning rate of 0.0001. The training takes advantage of early stopping where the patience is set to 5 epochs, with no improvements to the best validation error for more than 1 percent. The ratio of the validation set size to the available training data is set to 10%. The model is trained with 19 epochs on the training set (see Section 3.4) and takes about 6 hours on a GeForce RTX 3090. An inference pass in our setup requires on average less than 5 ms using the same setup.

Appendix C: Synthetic Data Generation Details

The 3DVHshadow generation details are provided in the following.

3D Models and Rigging. To synthesise shadows of people we use the 3DVH virtual human dataset [CMIH21] which contains 418 3D parametric models of people. These models are generated based on 14 male and 11 female bodies – with 8 to 48 modifications per body in shape and pose parameters, hair and clothing. 3DVH models are animated using the skeletal motion capture sequences from the Adobe Mixamo [Adob] database; in total, 50 different walking sequences are applied randomly to the parametric models. The clothing of people in 3DVH are from Adobe Fuse [Adoa]. In this work, we split the models into two sets of sizes 311 and 107 respectively for the training and evaluation.

Scene Contents. A scene contains a walking subject on a floor at a certain instant in time. A camera is placed at a random position looking at the subject's contact point on the floor. An isotropic point light source is positioned randomly in the scene.

Rendering Algorithm Settings. The Eevee rendering engine of Blender 3.0 [Ble] is employed to render the scene contents with the default settings, the *Standard* colour management transform mode and the target size of 512×512 resolution. The camera focal length is set to 20 mm and the sensor width to 36 mm. The soft shadow size of the point light source is set to 0 to produce hard shadows for training, and its shadow buffer bias and the clip start respectively to 0, and 0.25 metre. The shadow buffer settings are adjusted in order to render correct shadows at the contact points on the floor.

Factors of Variation. Each subject is assigned with a random posture and is rendered under 80 combinations of random point light and camera poses. Both the camera and light positions are sampled uniformly to have a radius of 2-5 metre to the subject (the hip position), and a height of 1.75-3.5 metre from the floor. This uniform sampling is repeated until there is even distribution of the shadow directions on circular sectors of the floor. In total, the dataset contains about 24400 training and 8400 test entries, leading to about 98K images. A random combination is discarded before rendering if it is known to produce a partial shadow on the ground plane in the image space. This is to avoid incomplete ground truth information.

Appendix D: Baseline Methods' Details

Baseline (b) Input Normalisation. To train the network, assuming that the subject is centred in the camera image, the subject's height is scaled to the height of the network input segmentation channel. The same scale is also applied to the corresponding camera image shadows. The inverse transforms are then applied on the network's output shadow. If the height-normalised shadows on the floor go outside the boundaries of the camera image, for fairness, they are ignored in our evaluation for all methods. The remaining 22686 training, and 7752 test entries in 3DVHshadow are therefore used in this experiment.

It is also worth noting that in the test images outside 3DVHshadow where the subject is not otherwise centred, this can be achieved by applying a homography transformation with a fixed centre of projection. This transformation has three degrees of freedom and can be calculated by having the subject's feet-floor contact point, and only the camera intrinsics. In a simplified case, the homography is KRK^{-1} where K is the camera intrinsics matrix and R the centring 3D rotation represented with Euler angles as $R_Z(\gamma)R_Y(\beta)R_X(\alpha)$. Assuming equal focal length in both directions, no sheer compensation, no sensor shift, and no camera roll ($\gamma = 0$), one can solve for α and β in a closed from manner (or by search) if focal length and the feet contact point are known or estimated.

PIFu's Training and Test. Training images are re-generated using PIFu's renderer with a weak-perspective camera model [HDL97] but according to the 2D camera angles and the source 3D objects of 3DVHshadow training set. To create the ground-truth occupancy function, we first centre all the 3D models to the origin and we sampled and labelled points on the surfaces following PIFu with $\mathcal{N}(0, \sigma = 0.05\text{m})$. We use RMSProp for the surface reconstruction following Newell et al. [NYD16], batch size of 4, learning rate 0.001. The learning rate of RMSProp is decayed by the factor of 0.1 at 10^{th} epoch. For the test phase, we employed 3DVHshadow test set directly, by cropping and aligning subjects to the image center and resizing the images to 512×512 as required by PIFu training specification. To calculate the metrics, the output estimated geometry is height normalised using the subject height and rotated to register to ground truth 3D mesh – both possible given the camera calibration and segmentation mask. The inference time for one input is about 5 seconds using an NVIDIA GeForce RTX 2070.