From Phase to Phenomenon: Self-Supervised Learning of Subsurface Scattering with Minimal Phase-shift Inputs

Arjun Majumdar, Raphael Braun, Andreas Engelhardt, and Hendrik PA. Lensch

Eberhard Karls Universität Tübingen, Germany
{arjun.majumdar,raphael.braun,andreas.engelhardt,
hendrik.lensch}@uni-tuebingen.de

Abstract. We propose a self-supervised pretraining framework for learning sub-surface scattering (SSS) light transport representations from minimal input. Our method leverages a stereo projector–camera setup that captures only eight high-frequency phase-shift profilometry (PSP) images per view to pretrain an encoder in a multi-view, multi-object setting. We introduce a tailored augmentation strategy for PSP-based SSS data, and show that it significantly outperforms standard ImageNet-style augmentations for SSL pretraining. The pretrained encoder learns generalizable SSS representations that transfer effectively to downstream tasks, including spatially varying relighting and representation evaluation using a kNN classifier. Combined with a decoder, the model reconstructs dense scattering footprint responses, trained using a dedicated cost function that improves accuracy, particularly for anisotropic footprints. An overview of our method is presented in Fig. 3. Despite using only eight input images per view, our approach generalizes to unseen objects with complex geometry and material properties, achieving high-fidelity reconstructions while requiring orders of magnitude fewer images than prior methods. Our code is publicly available at GitHub.

Keywords: Self-supervision · Subsurface Scattering · PSP

1 Introduction

Subsurface Scattering (SSS) occurs when light penetrates the surface of a material, scatters internally, and exits elsewhere. The scattering depth varies with the material and its geometry, effectively smoothing the incident light patterns. Analytical diffusion models describe this process and provide exact solutions for simple geometries [23]. Recent deep learning methods extend these solutions to complex surfaces while preserving efficiency [41]. Alternatively, volumetric path tracing can accurately simulate SSS but requires knowledge about the internal scattering properties of the material, which are difficult to obtain [9, 13]. Instead, we adopt an image-based acquisition and rendering approach that captures each light interaction as an image footprint. Relighting is then achieved

by combining these images with appropriate scaling. However, capturing the necessary dataset is expensive due to the non-local nature of SSS. We propose a novel, fully data-driven framework for image-based modeling of sub-surface scattering (SSS) that learns to accurately predict spatially varying SSS footprint responses at each surface point in a multi-view, multi-object setting. Once trained, the model enables scene relighting without requiring any additional acquisitions. Our method uses a non-contrastive self-supervised SimSiam objective [8] for pretraining an encoder backbone, which does not require explicit SSS footprint ground truth supervision. After pretraining, the learned representations are used to solve two downstream tasks: **Footprint reconstruction** A decoder is trained to predict dense SSS footprint responses across multiple views and across diverse objects. **kNN accuracy** The pretrained encoder learns generalizable SSS representations, whose quality is evaluated using a zero shot kNN classifier on the learned embeddings [43]. This setup demonstrates that the proposed self-supervised pretraining captures physically meaningful and generalizable SSS features that transfer effectively across tasks and objects.

The encoder is pre-trained in an SSL manner by using only 8 camera captured high-frequency phase shift profilometry (PSP) images as input for each object. PSP has been used for quite some time to capture accurate 3D geometry of scanned objects [49], [22]. Our encoder is trained on the same sinusoidal patterns that are used for projector-camera registration. Our setup generalizes to unseen views and objects with varying geometries and material properties. We conduct extensive qualitative and quantitative evaluations of the learned footprint responses by rendering objects under a black-white checkerboard virtual illumination pattern and compare the results against camera-captured ground truth. Since our approach focuses on learning and estimating each surface pixel footprint response, we emphasize detailed evaluations of the relit results.
Our contributions are as follows:

- We propose a data-efficient self-supervised framework for learning transferable subsurface scattering (SSS) representations from a small number of phase-shift profilometry (PSP) images per view. By pre-training the encoder, our method significantly reduces acquisition requirements compared to prior supervised approaches and enables effective reuse for downstream tasks.
- We present a physics-aware data augmentation pipeline specifically designed for high-frequency structured-light PSP inputs, enabling stable and effective self-supervised pretraining on SSS data.
- We design a task-specific loss formulation that accounts for the highly skewed and anisotropic distribution of SSS footprint responses. This tailored objective stabilizes training and improves reconstruction accuracy by explicitly encouraging the learning of anisotropic scattering behavior.
- We train a lightweight decoder to directly predict anisotropic pixel-level SSS footprint responses from encoded representations of projected high-frequency PSP patterns with explicit SSS footprint supervision. The encoder's learned representations are evaluated using a standard kNN classifier [43].

– We demonstrate strong generalization across unseen objects, geometries, materials, and viewpoints, validated through extensive qualitative and quantitative experiments.

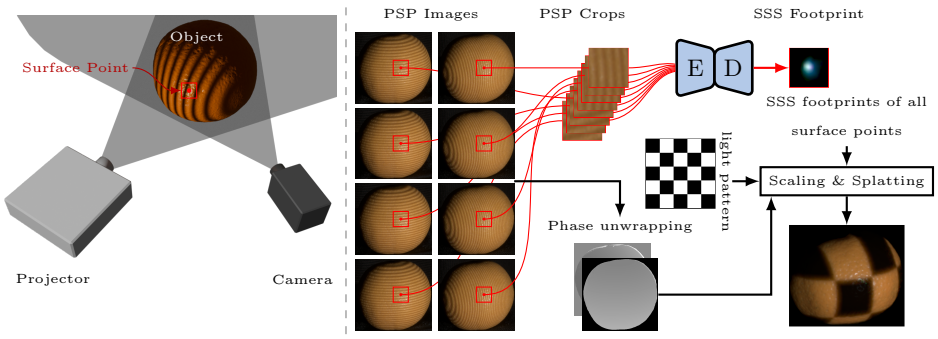


Fig. 1: Left: Our method works on an uncalibrated stereo projector-camera setup. **Center:** We capture images of the object under eight high-frequency sine wave patterns (PSP Images). Registration of camera- to projector-pixels is achieved via phase unwrapping. Our method estimates the SSS point spread function for any given surface point from a local crop of the PSP images. Relighting is achieved by scaling those SSS footprints for every surface point with the light they would receive from the desired light pattern, and splatting them onto a target canvas.

2 Background

Acquisition of material properties has a long history, including measurement of SSS parameters and general relightable representations such as reflectance fields [10]. Early methods directly sampled the impulse response to incident illumination, while later work employed fixed wavelet noise patterns [36], compressive sensing [37], and adaptive acquisition strategies [15, 34, 35, 39] that exploit the directional duality of light transport to accelerate the capture of general light transport matrices. Despite these advances, the number of required acquisition patterns remains large. In contrast, our method requires only eight input images per new object and view after an initial training stage across multiple objects to recover spatially varying pixel footprints for SSS.

Following the introduction of the first practical BSSRDF model [24], SSS parameters have been estimated using dedicated point-wise illumination setups [24, 42]. For complete objects, Goesele et al. [17] presented a system in which a single laser beam sequentially illuminates each surface point to measure the resulting footprint at the camera, requiring millions of individual measurements.

Recent neural relighting methods also target objects exhibiting SSS [11, 29, 44,46,47], but they typically assume distant illumination instead of incident light patterns and rely on one-light-at-a-time (OLAT) data for training.

Structured-light techniques, particularly those based on high-frequency patterns, have been used to separate reflected radiance components into local and global illumination effects [33]. Several 3D scanning approaches employ high-frequency PSP patterns for objects with SSS [6, 7, 14, 19, 32]. However, these methods primarily aim to suppress subsurface scattering artifacts to improve depth estimation rather than to measure or represent the scattering itself for relighting. Geiger et al. [16] also use structured light projection to correct topographic errors caused by volumetric light transport inside scattering materials, analyzing such effects through Monte Carlo simulations and introducing correction techniques based on quantified light propagation. Other studies [25, 26] leverage high-frequency binary patterns and polarization filters to directly measure local isotropic scattering parameters of human skin.

Vicini et al. [41] proposed a neural framework for efficiently sampling Bidirectional Surface Scattering Reflectance Distribution Functions (BSSRDFs) on complex 3D surfaces. Their system combines three components for conditioning a Conditional-VAE and MLP architecture. Majumdar et al. [30] present a U-Net [38] based method which leverages 3D scanning and a stereo projector-camera setup to learn pixel-level SSS responses for accurate, high-resolution relighting and generalization across materials and views. They use no data augmentation and/or SSL method and feed in raw inputs directly as inputs to map to SSS footprint responses for all surface points requiring expensive data acquisition and consequent processing. They train the encoder-decoder in an end-to-end manner. Therefore, their encoder cannot be used in a standalone manner for different tasks. Our method does not suffer from these shortcomings. To the best of our knowledge, we are the first to introduce and use data augmentation specifically tailored for high-frequency PSP SSS input patches for SSL non-contrastive pre-training. The pre-trained encoder is then used for the downstream task of reconstructing spatially varying, potentially anisotropic scattering footprints. Pre-training the encoder results in a simpler and more efficient training pipeline. The resulting neural model produces accurate, relightable representations of scattering behavior.

2.1 Phase-shifted Profilometry

Our acquisition of subsurface scattering representations builds on N-step Phase-Shifting Profilometry (PSP), a technique commonly used for high-accuracy 3D scanning [6, 7, 19, 22, 32, 48]. In a calibrated camera–projector setup, a sequence of N phase-shifted sinusoidal fringe patterns is projected onto the object and captured by the camera. Each projected pattern can be mathematically defined as

$$I^p(x^p, y^p) = a^p + b^p \cos(2\pi f_0^p x^p + \delta_n), \tag{1}$$

where (x^p, y^p) are projector coordinates, a^p is the average intensity, b^p the amplitude, f_0^p the fringe frequency, and $\delta_n = 2\pi n/N$ the phase shift.

(a) Apple **(b)** Pear **(c)** Star

Fig. 2: Effect of SSS on PSP images. Strongly scattering materials (apple) blur projected patterns and reduce contrast due to wider scattering footprints, whereas weakly scattering objects (pear, star) preserve sharper sinusoidal structures.

The corresponding camera captured images are modeled as

$$I_n(x,y) = A(x,y) + B(x,y)\cos\left(\phi(x,y) - \frac{2\pi n}{N}\right), \tag{2}$$

where $A(x,y)$ denotes background illumination, or indirectly scattered illumination, $B(x,y)$ the modulation (linked to reflectivity), and $\phi(x,y)$ the wrapped phase, computed as

$$\phi(x,y) = \tan^{-1}\frac{\sum_{n=0}^{N-1} I_n(x,y)\sin(2\pi n/N)}{\sum_{n=0}^{N-1} I_n(x,y)\cos(2\pi n/N)}. \tag{3}$$

The wrapped phase $\phi(x,y) \in [-\pi,\pi]$ is converted to an unwrapped phase $\Phi(x,y) = \phi(x,y) + 2\pi k(x,y)$, where $k(x,y)$ indicates fringe order, correlating directly with scene depth in stereo geometry.

Phase unwrapping, i.e. determining global offsets based on locally recovered phases $\phi(x,y)$, can be performed spatially using neighboring pixels or temporally using phase shifts over time. Temporal methods are generally more robust, especially near surface discontinuities [22]. However, for translucent or highly scattering materials, traditional PSP often fails because subsurface scattering distorts the low-frequency phase maps required for reliable unwrapping [6, 7]. We capture eight images for $N = 4$ step PSP: four with height-varying and four with width-varying sinusoidal patterns. Example unwrapped phase maps for the orange is depicted in Fig. 1.

SSS affects captured PSP images by spatially spreading the reflected light beneath the surface. Materials with stronger scattering blur the projected sinusoidal patterns and reduce their local contrast, which will be encoded into the latent representations. For example, highly scattering objects such as apples exhibit significantly attenuated and smoother PSP patterns, whereas objects with weaker scattering, such as pears, preserve sharper pattern contrasts which is shown in Fig. 2.

3 Method

We present an overview of our method in Fig. 3 which on LHS shows SSL pretraining of the encoder, while RHS shows the frozen encoder used for the two downstream tasks, viz., training the decoder on SSS footprint responses and kNN zero-shot classification.

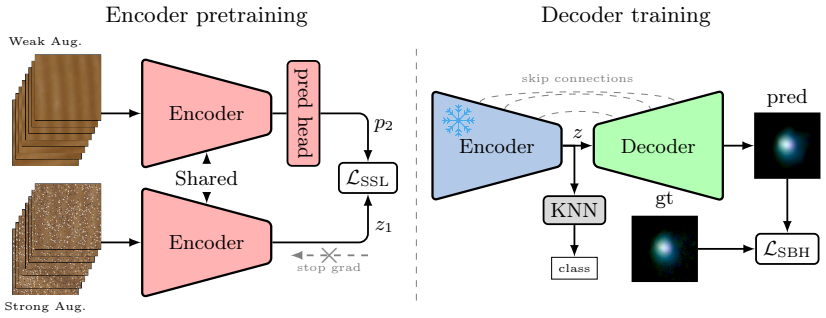


Fig. 3: Overview of proposed framework. Left: SSL pretraining of shared encoder using 2 augmented PSP patches, minimizing L_{SSL}. **Right:** The decoder is trained to predict SSS footprints in a supervised setting, exploiting the embedding space of the frozen encoder. It is optimized based on ground truth SSS footprints and an auxiliary classification task. Classification is done with a zero-shot kNN approach directly on the encoder features.

3.1 Self-Supervised Learning

Self-Supervised Learning (SSL) in computer vision aims to learn transferable and generalizable feature representations without relying on manual annotations. Most modern SSL approaches are based on similarity learning, where a backbone network f is trained to map semantically related inputs to nearby points in a high-dimensional embedding space. Typically, these related inputs are simulated during training as different augmentations of the same image.

Many contemporary methods adopt a Siamese architecture consisting of weight-sharing neural network branches that process paired views of an image. Although this framework enables effective representation learning, it also introduces the risk of *representation collapse*, where the network converges to a degenerate solution by mapping all inputs to an almost constant embedding.

Extensive prior work has analyzed this collapse phenomenon and proposed mechanisms to prevent it. In general, SSL methods can be categorized into three groups: **Contrastive methods:** These approaches explicitly pull positive (similar) embeddings closer while pushing negative (dissimilar) embeddings apart. The use of negative samples prevents trivial constant mappings and stabilizes

training [4], [20], [2]. **Non-contrastive methods:** These methods align positive pairs without relying on negative samples. To avoid collapse, architectural asymmetries (e.g., predictor networks, stop-gradient operations) or implicit constraints are introduced [18], [8]. **Regularization-based methods:** These approaches impose explicit statistical constraints on the embedding space, such as encouraging feature variance, decorrelation across embedding dimensions, or enforcing the cross-correlation matrix to approximate identity [45], [1], [12]. Such constraints ensure diverse and non-degenerate representations. This taxonomy provides a unified perspective on how modern SSL frameworks prevent collapse while learning meaningful visual representations.

We propose a self-supervised pretraining framework for learning sub-surface scattering (SSS) light transport representations from high-frequency phase-shift profilometry (PSP) images. We pretrain an encoder in a multi-view, multi-object setting to learn a generalizable SSS representation. We then train a decoder to predict the SSS responses and a prediction head to estimate the material class from those representations in a supervised manner. Figure 3 visualizes our training pipeline for both stages.

3.2 Data Acquisition

For the encoder, we capture 8 PSP images for every object. For training the decoder, we further capture the ground truth scattering footprints by illuminating individual dots on the surface for several training objects. We use a stereo projector-camera setup to capture both pixel-level SSS responses (point spread functions PSF) and the PSP images as shown in Fig. 1. The acquisition setup and basic image processing follow [30]. The processing pipeline for all captured images includes the acquisition of raw Bayer images, median filtering each RGGB channel with a 3×3 filter, demosaicing with Malvar's algorithm [31] and finally masking the objects using SAM [27]. Further details are in supplemental (S1).

For the encoder, we use high-frequency PSP patterns with 4 shifts. This means that we only capture 8 images per-view per object, 4 with vertical and 4 with horizontal orientation. No more than these 8 images per view are are necessary to train the encoder and to do the final inference of the scattering response footprints. We crop patches of size (90×90) pixels from those 8 images and provide them as input to the encoder as indicated in Fig. 1 and Fig. 3.

Local SSS footprint responses are captured by projecting dot patterns that illuminate individual points on the object's surface. This grid pattern enables the parallel acquisition of SSS responses for multiple surface points in one image. The distance between each illuminated pixel in projector space is chosen such that the footprints do not overlap (55 pixels in both x and y directions), The result of this acquisition can be seen on the right side of Figure 4. The grid pattern is shifted that every surface point is illuminated once. This acquisition is repeated for multiple views per object and for multiple different *training* objects. The total number of such SSS footprint response samples = 1126827 in total for 5 objects (Green apple, Orange, Pear, Star and Shovel) with up to four views each. For inference on novel objects, the trained encoder-decoder network can

directly predict the footprints purely based on the 8 PSP images, not requiring any captured footprints.

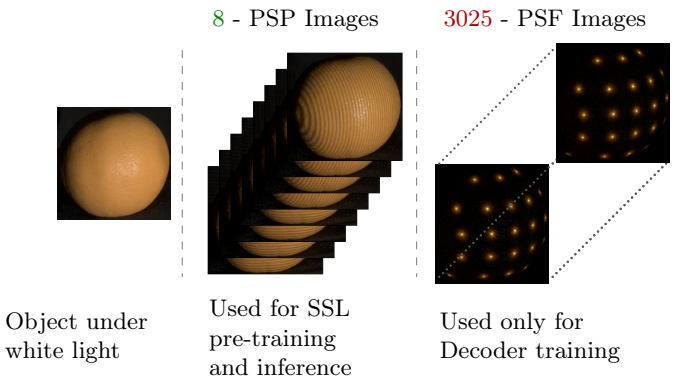


Fig. 4: For pre-training the Encoder with SSL we only use 8 PSP images per object. The point spread functions (PSF) of every surface point are captured in parallel with a spacing of 55 pixels, which requires 3025 images. Those images are only captured and used to train the Decoder once. The decoder generalizes to unseen objects during inference. Thus for relighting new objects we only have to capture 8 PSP images.

3.3 SSL Pretraining

Typical SSL methods use either ResNet-50 or ViT backbones [1, 4, 5, 18, 20, 28]. We do not adopt these architectures since our framework additionally requires a decoder to learn SSS footprint responses. A vanilla U-Net block [conv $\rightarrow$ BN $\rightarrow$ ReLU] $\times 2$ is less effective than residual blocks for learning stable embeddings and is more prone to dimensional collapse. Therefore, we design a custom architecture inspired by pre-activation residual blocks [21]. Detailed encoder and decoder architectures are provided in supplemental (S2). We pre-train the encoder with SimSiam's [8] non-contrastive method to learn generic representations of the high-frequency PSP input patches for each point on the object's surface. We provide ablations of using SimSiam's hyper-parameters vs. our modified ones together with dimensional collapse plots in the supplemental. During training, the sets of PSP patches are randomly augmented to two different views (Sec. 3.4). The training loss forces the output latent vectors to be invariant to the augmentations (see Fig. 3).

The crops for pretraining are extracted from every point on the object's surface within the SAM mask. For each point, a $(90, 90)$ pixel patch around the pixel is cropped out for all 8 PSP images. They are used as input for minimizing the negative cosine similarity on normalized vectors for the SimSiam method [8],

Fig. 5: Demonstration of our PSP specific strong augmentations for one input patch.

which can be mathematically written as:

$$\mathcal{L}_{\text{SSL}}(p_1, z_2) = -\frac{p_1}{\|p_1\|_2} \cdot \frac{z_2}{\|z_2\|_2}. \tag{4}$$

3.4 Augmentation

One of the most important factors in SSL pre-training is the employed data augmentation. There are standard data augmentation pipelines that are tailored for natural images. We have very specific types of images, namely PSP SSS patches, for which we developed our own specialized data augmentation pipeline.

Following [28], instead of strong-strong augmentations on both network branches, we apply a weak-strong augmentation pipeline to stabilize training and to obtain better generalizing representations. Consequently, we formulate our custom data augmentations for high-frequency PSP input patches as: **Weak augmentation:** We follow [28] with random-resized cropping and slight color jitter. **Our Strong augmentation:** salt&pepper noise, Gaussian blur, horizontal or vertical flips on all PSP images, random intensity scaling, random phase-jitter (add Gaussian (phase-like) noise to RGB images) and randomly shuffled PSP images (to break fixed $\frac{2\pi}{N}$ phase-shifts symmetry between the camera images). The effect of our new strong augmentations on a patch is visualized in Fig. 5.
In principle, the trained representations z could be useful for any downstream task. We specifically use them to train a robust footprint prediction decoder, which generalizes to unseen objects.

3.5 Supervised Decoder Training

To estimate the SSS footprints from PSP patches, the encoder is kept frozen while the decoder is trained in a supervised fashion, as shown on the right side of Fig. 3. For each pixel, we crop a 90×90 patch from the PSF images centered at that pixel and encode it into a representation z. The decoder then predicts the local SSS response, supervised by camera-captured ground truth SSS responses for surface points on the object using our Self-Balanced Hybrid Loss $\mathcal{L}_{\text{SBH}}$ (Sec. 3.6).

In the decoder, each decoder stage performs progressive spatial upsampling with attention-gated skip fusion using sub-pixel upsampling [40]. This formulation can be interpreted as a *machine translation* problem, where the network

maps PSP patch observations to a different output domain corresponding to local SSS footprint responses. Training the decoder object-by-object leads to catastrophic forgetting; therefore, we randomly sample patches across all objects and views during training.

Relighting Pipeline Let L denote the virtual projector image and π the camera–projector pixel correspondence recovered from the unwrapped phase. For each surface pixel p: (i) encode PSP crop 90×90 centered on p and decode to obtain predicted SSS footprint $\mathbf{PSF}_p \in \mathbb{R}^{90\times90\times3}$; (ii) look up projector color $c(p) = L(\pi(p)) \rightarrow$ this is the **scaling** step; (iii) **splat** each scaled footprint into relit canvas and accumulate: $I(q) = \sum_p \mathbf{PSF}_p(q-p)\cdot c(p)$, where $c(p)$ scales the contribution of light entering at p and the sum over p performs the splat-and-accumulate.

3.6 Self-Balanced Hybrid Loss

The light footprints caused by subsurface scattering typically have a strong peak in the point that is illuminated with an exponential radial falloff. We designed a physics-aware hybrid loss tailored for reconstructing those footprints. Our loss consists of five components:

$$\mathcal{L}_{\text{SBH}} = \lambda_1\mathcal{L}_1 + \lambda_{\log}\mathcal{L}_{\log} + \lambda_{\text{SSIM}}\mathcal{L}_{\text{SSIM}} + \lambda_\nabla\mathcal{L}_\nabla + \lambda_{\text{wMSE}}\mathcal{L}_{\text{wMSE}}.$$

$\mathcal{L}_1$: Standard L1 loss as data term. Works well on low-intensity data and prevents over-smoothing.

$\mathcal{L}_{\log}$: L1 loss on logarithmic data. This increases sensitivity to errors in faint regions.

$\mathcal{L}_{\text{SSIM}}$: Structural similarity loss as perceptual component.

$\mathcal{L}_\nabla$: L1 loss on channel normalized gradients. The normalization of the gradient of each channel prevents color-scale collapse.

$\mathcal{L}_{\text{wMSE}}$: Pixel-wise variance across the batch is used to weight the loss, placing greater emphasis on high-variance regions of the SSS footprint by using the relative variance as weight: $\sigma^2_{cij} = \text{Var}_b(Y^{(b)}_{cij})$ and $w_{cij} = 1 + \sigma^2_{cij}/(\bar{\sigma}^2 + \varepsilon)$. We use this pixel wise weight w_{cij} in a standard MSE loss.

The proposed objective jointly sharpens the peaks ($\mathcal{L}_\nabla$, $\mathcal{L}_{\text{wMSE}}$), preserves SSS tails ($\mathcal{L}_{\log}$), stabilizes global intensity ($\mathcal{L}_1$), and maintains perceptual structure ($\mathcal{L}_{\text{SSIM}}$). Batch-adaptive variance weighting enables self-balancing behavior, mitigating the intrinsic smoothing bias of convolutional networks. In our experiments, we set the loss weights to $\lambda_1 = 1$, $\lambda_{\text{WMSE}} = 5$, $\lambda_\nabla = 10$, $\lambda_{\text{SSIM}} = 0.5$, $\lambda_{\log} = 1$.

3.7 Relighting

Given our trained encoder and decoder, we can relight any object given eight sinusoidal fringe images. For every surface point, we crop 90×90 patches around

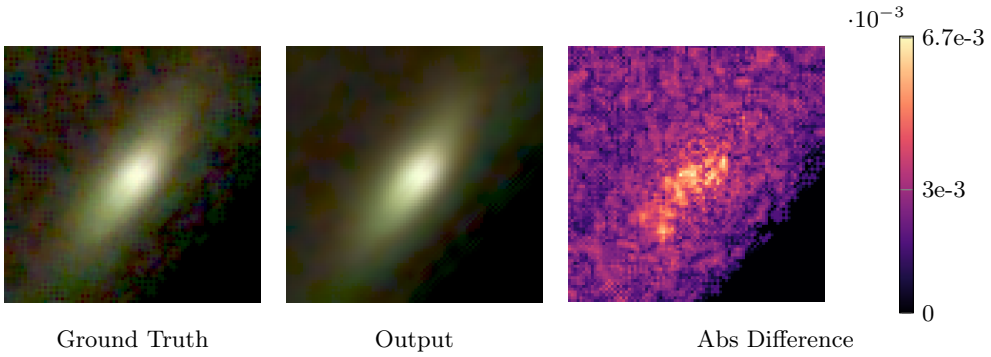


Fig. 6: Qualitative comparison of SSS footprints. Note the denoised quality of the predicted strongly anisotropic footprint.

the point from the PSP images, obtain their latent representation with the encoder and finally decode the latent to obtain the pixel's SSS response footprint. Relighting then boils down to scaling those responses with the light received from the virtual projector image and splatting the scaled footprint onto the image canvas. This is visualized on the right side of Fig. 1.

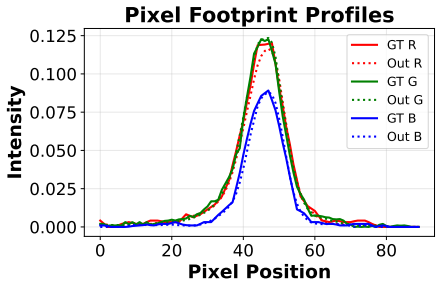
4 Results

4.1 Training Details

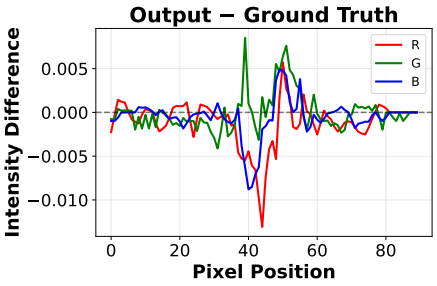
When using SimSiam's hyperparameters, we observe a dimensional collapse due to the different data distribution. Therefore, we modify the parameters and start with an initial lr=0.05 *without scaling* accompanied with lr warmup (20 warmup epochs), followed by cosine decay without warm restarts. This modification allows us to get a stable training without dimensional collapse. Training details, loss curve visualizations, dimensional collapse and further analysis are in the supplemental.

4.2 Assessment of SSS Footprint Reconstructions

A qualitative comparison of the camera-captured GT footprint responses vs. their corresponding output responses, along with L1-difference aggregated for RGB channels, is shown in Fig. 6. Each footprint response patch is $(90, 90, 3)$ pixels. The trained encoder-decoder network accurately reconstructs the responses even for strongly anisotropic footprints. The apparent camera noise of the GT capture is significantly reduced. To also visualize the quantitative behavior, Fig. 7 shows the intensity profile across the same footprint as in Fig. 6. The reconstruction accurately follows the ground truth for the three color channels with minor deviations in the peak and precise shape in the extended tail. For more visualizations and reconstructed footprints see the supplemental. Aggregated reconstruction metrics for unseen test objects are reported in Tab. 1.



(a) Pixel footprint profiles. Solid lines denote ground truth, dotted lines show the network prediction.



(b) Difference between prediction and ground truth profiles.

Fig. 7: Comparison: predicted vs. ground-truth SSS footprint responses, showing one scan line for a representative patch. Note the faithful prediction from the PSP input even in the tail of the scattering profile.

Table 1: Aggregated quantitative reconstruction metrics for different test objects and views (MSE$\times 10^{-5}$, LPIPS$\times 10^{-2}$).

Object	MSE↓	PSNR↑	SSIM↑	LPIPS↓
Apple	4.1	44.67	0.96	8.4
Orange	4.9	43.84	0.94	8.6
Crab	3.5	47.98	0.98	6.5
Hand	3.8	42.51	0.96	6.5

Table 2: kNN accuracy of PSP-specific augmentations show improved embedding quality compared to ImageNet ones.

Object	PSP↑	ImageNet↑
Apple	**99.6**	99.04
Pear	**99.99**	98.74
Orange	**99.99**	98.92
Star	**99.95**	99.93
Shovel	**91.43**	84.91

4.3 Relighting Results

For relighting, we first reconstruct the SSS response footprints for every individual pixel and then can use any virtual projector image matching the original projector's resolution (1920x1080). Qualitative relighting results on unseen objects are shown in Fig. 8. The precise phase unwrapping ensures perfectly aligned incident light patterns. The trained network faithfully reconstructs the objects' scattering appearance, in particular, the edges between lit and unlit areas, where SSS leads to smoothed transitions. On the tangerine, even the highlights are properly shown. In the toy sand rake example, the scattering and interreflections between the complex geometry of the fingers are correctly reproduced.

Relighting results for multiple views of the test objects *Apple* and *Crab* (both unseen during training) are shown in Fig. 9. The replicated scattering behavior is of consistent quality across the views. Besides the accurate surface texture, note the subtle color variation in the indirectly lit dark checkers of the red and the yellow side of the same apple, indicating spatially varying scattering. In the crab example, our approach correctly makes some geometric detail visible in the dark

fields. We demonstrate further generalization by relighting two **new unseen** objects — a leaf (heterogeneous, fine-veined) and a LEGO brick (anisotropic, edged) — under **a novel stripe pattern** in Fig. 10.

Our combined SSL and supervised training scheme reproduces even subtle scattering behavior while drastically reducing the acquisition requirements for any new object: While DISCO [17], based on individual laser beams, requires millions of images and [30] requires about 3000 images for novel objects (shifted pixel grids), our approach requires only 8 PSP input images per view for performing accurate inference even on unseen objects. Refer supplemental for more details.

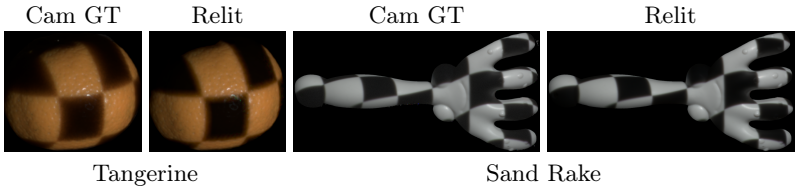


Fig. 8: Qualitative comparisons on two test objects. Cam GT is a real photo of the object lit by the projector, Relit is the result of our relighting. Note: Those objects have never been seen during training of the encoder or decoder.

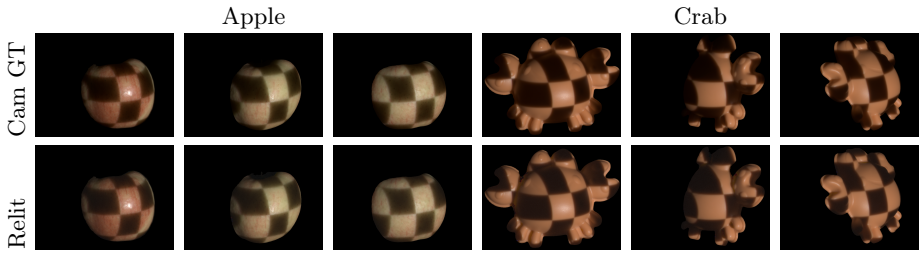


Fig. 9: Multiple relit views of two objects using the per-pixel reconstructed SSS footprints compared to camera-captured ground truth, where the real projector illuminates the scene with the same pattern. Note those objects have never been seen during training of the encoder or decoder.

4.4 kNN Representation Evaluation

To evaluate the quality of the learned patch embeddings, we perform a k-nearest neighbor (kNN) classification experiment using frozen encoder features [43]. Each object in the dataset is captured from V viewpoints. To avoid data leakage from

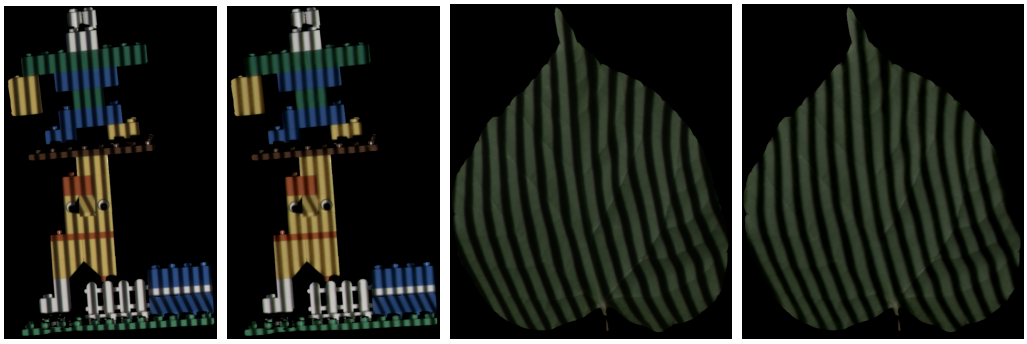


Fig. 10: LEGO & Leaf relit under stripe illumination on unseen objects and illumination.

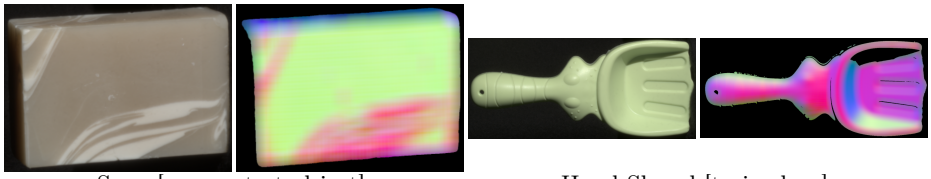
spatially adjacent patches within the same image, we adopt a *view-level leave-one-out* protocol: all patches from one view are held out for evaluation, while patches from the remaining views form the feature database.

For a query patch, its embedding is compared to the database using cosine similarity, and the label is predicted via majority voting among the k nearest neighbors. We report two complementary metrics: (1) **overall accuracy**, defined as the fraction of correctly classified patches across all held-out views, and (2) **mean per-view accuracy**, computed by averaging classification accuracy across views, thereby giving equal weight to each viewpoint.

This evaluation protocol is entirely training-free: the encoder remains frozen and no fine-tuning or linear probing is performed. Consequently, strong kNN performance indicates that the learned representations capture discriminative material characteristics. Quantitative results comparing PSP-specific augmentations with standard ImageNet augmentations are reported in Table 2. consistently achieve higher accuracy. A detailed analysis is in supplemental (6.2).

4.5 Visualization via PCA-RGB Feature Mapping

To visualize the spatial distribution of SSS properties learned by the pretrained encoder, we map patch features to colors using a PCA-RGB projection inspired by [3]. Since the encoder is trained via SimSiam to produce consistent representations for patches with similar scattering behavior, spatially similar SSS properties cluster in feature space. PCA preserves this structure, so similar colors correspond to similar subsurface scattering characteristics, while color discontinuities indicate material boundaries or changes in scattering depth. Examples are shown in Fig. 11. For an object with nearly homogeneous scattering, the visualization resembles surface normal maps due to the dependence on local illumination directions, whereas spatially varying materials exhibit clear color variation. Additional examples and details are provided in Supplemental (6.1).



Soap [unseen test object] Hand Shovel [trained on]

Fig. 11: PCA-RGB visualization of the learned SSS features for different objects. On the soap with heterogeneous materials, the visualization clearly marks the veins. On the hand shovel, the normal directions can be seen.

4.6 Limitations and Future Work

Our method reconstructs accurate SSS footprints for the specific camera–projector configuration used during acquisition. Although it generalizes across views and geometric changes, footprints remain dependent on local illumination and viewing directions, and interpolation across camera or projector positions remains future work. Moreover, scattering is restricted to the local (90×90) patches, limiting the recoverable scattering radius and preventing longer-range transport. Finally, reliance on standard dynamic range images may introduce noise and dynamic-range limitations in the recovered SSS responses.

Future work could explore HDR acquisitions to improve signal quality and investigate generative or stochastic modeling approaches for SSS footprints instead of deterministic regression, potentially enabling better representation of complex or highly anisotropic scattering behavior.

5 Conclusion

By using only eight high-frequency PSP images, our model learns to predict accurate, anisotropic scattering footprints that generalize to novel objects and viewpoints. This was enabled by splitting the encoder-decoder translation into two separate training tasks: stable self-supervised training of the encoder, which only required easy-to-acquire PSP images to produce a versatile latent representation, followed by supervised training of the decoder to predict accurate scattering impulse-response footprints. We showed that domain-specific augmentations and a tailored hybrid loss are critical for maintaining structural and radiometric fidelity. Our method drastically reduces the number of images needed for precise footprint predictions to only 8 images per view.

Acknowledgements

This work has been supported by the Deutsche Forschungsgemeinschaft (DFG) – EXC number 2064/1 – Project number 390727645 and SFB 1233, TP 2, Project number 276693517, and by the International Max Planck Research School for Intelligent Systems (IMPRS-IS).

References

1. Bardes, A., Ponce, J., LeCun, Y.: Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In: ICLR. OpenReview.net (2022)
2. Caron, M., Misra, I., Mairal, J., Goyal, P., Bojanowski, P., Joulin, A.: Unsupervised learning of visual features by contrasting cluster assignments. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H.T. (eds.) NeurIPS (2020)
3. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: ICCV (2021)
4. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.E.: A simple framework for contrastive learning of visual representations. CoRR **abs/2002.05709** (2020)
5. Chen, T., Kornblith, S., Swersky, K., Norouzi, M., Hinton, G.E.: Big self-supervised models are strong semi-supervised learners. CoRR **abs/2006.10029** (2020)
6. Chen, T., Lensch, H.P.A., Fuchs, C., Seidel, H.P.: Polarization and phase-shifting for 3d scanning of translucent objects. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2007). https://doi.org/10.1109/CVPR.2007.383209
7. Chen, T., Seidel, H.P., Lensch, H.P.A.: Modulated phase-shifting for 3d scanning. In: 2008 IEEE Conference on Computer Vision and Pattern Recognition. pp. 1–8 (2008). https://doi.org/10.1109/CVPR.2008.4587836
8. Chen, X., He, K.: Exploring simple siamese representation learning. In: CVPR. pp. 15750–15758. Computer Vision Foundation / IEEE (2021)
9. Christensen, P., Fong, J., Shade, J., Wooten, W., Schubert, B., Kensler, A., Friedman, S., Kilpatrick, C., Ramshaw, C., Bannister, M., Rayner, B., Brouillat, J., Liani, M.: Renderman: An advanced path-tracing architecture for movie rendering. ACM TOG **37** (2018)
10. Debevec, P., Hawkins, T., Tchou, C., Duiker, H.P., Sarokin, W., Sagar, M.: Acquiring the reflectance field of a human face. In: Proceedings of the 27th annual conference on Computer graphics and interactive techniques. pp. 145–156 (2000)
11. Dihlmann, J.N., Majumdar, A., Engelhardt, A., Braun, R., Lensch, H.: Subsurface scattering for gaussian splatting. In: The Thirty-eighth Annual Conference on Neural Information Processing Systems (2025)
12. Ermolov, A., Siarohin, A., Sangineto, E., Sebe, N.: Whitening for self-supervised representation learning. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 3015–3024. PMLR (18–24 Jul 2021)
13. Fascione, L., Hanika, J., Leone, M., Droske, M., Schwarzhaupt, J., Davidovič, T., Weidlich, A., Meng, J.: Manuka: A batch-shading architecture for spectral path tracing in movie production. ACM TOG **37** (2018)
14. Fuchs, C., Heinz, M., Levoy, M., Seidel, H.P., Lensch, H.P.: Combining confocal imaging and descattering. In: Computer Graphics Forum. vol. 27, pp. 1245–1253. Wiley Online Library (2008)
15. Garg, G., Talvala, E.V., Levoy, M., Lensch, H.P.A.: Symmetric photography: Exploiting data-sparseness in reflectance fields. In: Proceedings of the Eurographics Symposium on Rendering. pp. 251–262. Eurographics Association (2006)
16. Geiger, S., Hank, P., Kienle, A.: Improved topography reconstruction of volume scattering objects using structured light. Journal of the Optical Society of America A **39** (2022)
17. Goesele, M., Lensch, H.P., Lang, J., Fuchs, C., Seidel, H.P.: Disco: acquisition of translucent objects. In: ACM SIGGRAPH 2004 Papers, pp. 835–844. ACM (2004)

18. Grill, J.B., Strub, F., Altché, F., Tallec, C., Richemond, P.H., Buchatskaya, E., Doersch, C., Pires, B.Á., Guo, Z., Azar, M.G., Piot, B., Kavukcuoglu, K., Munos, R., Valko, M.: Bootstrap your own latent - a new approach to self-supervised learning. In: Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.F., Lin, H.T. (eds.) NeurIPS (2020)
19. Gupta, M., Nayar, S.K.: Micro phase shifting. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition. pp. 813–820 (2012). https://doi.org/10.1109/CVPR.2012.6247753
20. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.B.: Momentum contrast for unsupervised visual representation learning. In: CVPR. pp. 9726–9735. Computer Vision Foundation / IEEE (2020)
21. He, K., Zhang, X., Ren, S., Sun, J.: Identity mappings in deep residual networks. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) Computer Vision – ECCV 2016. pp. 630–645. Springer International Publishing, Cham (2016)
22. He, X., Kemao, Q.: A comparative study on temporal phase unwrapping methods in high-speed fringe projection profilometry. Optics and Lasers in Engineering **142** (2021). https://doi.org/10.1016/j.optlaseng.2021.106613
23. Jensen, H.W., Marschner, S.R., Levoy, M., Hanrahan, P.: A practical model for subsurface light transport. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques. pp. 511–518. SIGGRAPH '01, Association for Computing Machinery, New York, NY, USA (2001). https://doi.org/10.1145/383259.383319
24. Jensen, H.W., Marschner, S.R., Levoy, M., Hanrahan, P.: A practical model for subsurface light transport. In: Proceedings of the 28th annual conference on Computer graphics and interactive techniques. pp. 511–518 (2001)
25. Kienle, A., Lilge, L., Patterson, M.S., Hibst, R., Steiner, R., Wilson, B.C.: Spatially resolved absolute diffuse reflectance measurements for noninvasive determination of the optical scattering and absorption coefficients of biological tissue. Applied optics **35**(13), 2304–2314 (1996)
26. Kikuchi, K., Katsuyama, M., Shibata, T., Hardeberg, J., Yuasa, T., Aizu, Y.: Development of measurement system for subsurface scattering light of skin and analysis of its age-related changes. In: Biomedical Imaging and Sensing Conference. vol. 12608, pp. 123–126. SPIE (2023)
27. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.: Segment anything. 2023 IEEE/CVF International Conference on Computer Vision (ICCV) (2023). https://doi.org/10.1109/iccv51070.2023.00371
28. Koohpayegani, S.A., Tejankar, A., Pirsiavash, H.: Mean shift for self-supervised learning. In: ICCV. pp. 10306–10315. IEEE (2021)
29. Lyu, L., Tewari, A., Leimkühler, T., Habermann, M., Theobalt, C.: Neural radiance transfer fields for relightable novel-view synthesis with global illumination. In: European Conference on Computer Vision. pp. 153–169. Springer (2022)
30. Majumdar, A., Braun, R., Lensch, H.: Neural Acquisition & Representation of Subsurface Scattering. In: Egger, B., Günther, T. (eds.) Vision, Modeling, and Visualization. The Eurographics Association (2025). https://doi.org/10.2312/vmv.20251228
31. Malvar, H.S., wei He, L., Cutler, R.: High-quality linear interpolation for demosaicing of Bayer-patterned color images. In: Proceedings of the 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). vol. 4, pp. iv–485–8 (May 2004). https://doi.org/10.1109/ICASSP.2004.1326587

32. Moreno, D., Son, K., Taubin, G.: Embedded phase shifting: Robust phase shifting with embedded signals. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 2301–2309 (2015). https://doi.org/10.1109/CVPR.2015.7298843
33. Nayar, S., Krishnan, G., Grossberg, M., Raskar, R.: Fast separation of direct and global components of a scene using high frequency illumination. ACM Trans. Graph. **25**, 935–944 (Jul 2006). https://doi.org/10.1145/1179352.1141977
34. O'Toole, M., Kutulakos, K.N.: Optical computing for fast light transport analysis. ACM Trans. Graph. **29**(6), 164 (2010)
35. O'Toole, M., Raskar, R., Kutulakos, K.N.: Primal-dual coding to probe light transport. ACM Trans. Graph. **31**(4), 39–1 (2012)
36. Peers, P., Dutré, P.: Wavelet environment matting. In: Proceedings of the 14th Eurographics workshop on Rendering. pp. 157–166 (2003)
37. Peers, P., Mahajan, D.K., Lamond, B., Ghosh, A., Matusik, W., Ramamoorthi, R., Debevec, P.: Compressive light transport sensing. ACM Transactions on Graphics (ToG) **28**(1), 1–18 (2009)
38. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. CoRR **abs/1505.04597** (2015)
39. Sen, P., Chen, B., Garg, G., Marschner, S., Sen, M., Zoras, P., Lunsford, P., Genetti, J., Levoy, M.: Dual photography. ACM Transactions on Graphics **24**(3), 745–755 (2005). https://doi.org/10.1145/1073204.1073257
40. Shi, W., Caballero, J., Huszár, F., Totz, J., Aitken, A.P., Bishop, R., Rueckert, D., Wang, Z.: Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1874–1883 (2016). https://doi.org/10.1109/CVPR.2016.207
41. Vicini, D., Koltun, V., Jakob, W.: A learned shape-adaptive subsurface scattering model. ACM TOG **38** (2019)
42. Weyrich, T., Matusik, W., Pfister, H., Bickel, B., Donner, C., Tu, C., McAndless, J., Lee, J., Ngan, A., Jensen, H.W., et al.: Analysis of human faces using a measurement-based skin reflectance model. ACM Transactions on Graphics (ToG) **25**(3), 1013–1024 (2006)
43. Wu, Z., Xiong, Y., Yu, S.X., Lin, D.: Unsupervised feature learning via non-parametric instance discrimination. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3733–3742 (2018). https://doi.org/10.1109/CVPR.2018.00393
44. Yu, H.X., Guo, M., Fathi, A., Chang, Y.Y., Chan, E., Gao, R., Funkhouser, T., Wu, J.: Learning object-centric neural scattering functions for free-viewpoint relighting and scene composition. Trans. Mach. Learn. Res. (2023). https://doi.org/10.48550/arXiv.2303.06138
45. Zbontar, J., Jing, L., Misra, I., LeCun, Y., Deny, S.: Barlow twins: Self-supervised learning via redundancy reduction. In: Meila, M., Zhang, T. (eds.) ICML. Proceedings of Machine Learning Research, vol. 139, pp. 12310–12320. PMLR (2021)
46. Zheng, Q., Singh, G., Seidel, H.P.: Neural relightable participating media rendering. Advances in Neural Information Processing Systems **34**, 15203–15215 (2021)
47. Zhu, S., Saito, S., Bozic, A., Aliaga, C., Darrell, T., Lassner, C.: Neural relighting with subsurface scattering by learning the radiance transfer gradient. arXiv preprint arXiv:2306.09322 (2023)
48. Zuo, C., Feng, S., Huang, L., Tao, T., Yin, W., Chen, Q.: Phase shifting algorithms for fringe projection profilometry: A review. Optics and Lasers in Engineering **109** (2018). https://doi.org/10.1016/j.optlaseng.2018.04.019

49. Zuo, C., Huang, L., Zhang, M., Chen, Q., Asundi, A.: Temporal phase unwrapping algorithms for fringe projection profilometry: A comparative review. Optics and Lasers in Engineering **85** (2016). https://doi.org/10.1016/j.optlaseng.2016.04.022