A Design of Contractive Appearance Flow for Photometric Stereo

Lixiong Chen
lchen@robots.ox.ac.uk
Victor Adrian Prisacariu
victor@robots.ox.ac.uk
Active Vision Lab
University of Oxford
Oxford, UK

Abstract

We introduce the concept of contractive appearance flow to address photometric stereo with general reflectance. Our solution is motivated by the fact that the shape intrinsics of an object are encoded by its Lambertian reflectance, based on which we design a neural network that maps a set of per-pixel general appearances to their Lambertian counterparts as if this process is carried by a flow in a field of vectors of pixel values. Our design has two features: (1) by introducing a transfer operator in the encoded latent space, we replace the typical workflow of a Variational AutoEncoder (VAE) with a more generic encode-transfer-decode procedure. For photometric stereo, we apply this procedure to produce consistent representations of the incident light fields and to eliminate the signal variation caused by material properties; (2) during training each sample of general reflectance is associated with its Lambertian-related template samples, and by minimizing the distance between these two types of signals in the latent space, we enforce the flow to contract in the subspace spanned by Lambertian appearances only. The proposed method learns reflectance measurements directly and does not need to parameterize material properties. Our design is simple, lightweight, and automatic, yet, experiments show that it is effective and yields accurate estimations.

1 Introduction

An effective inverse model for surface reflectance plays an important role in data-driven photometric stereo [51]. A common paradigm is to establish a mapping from reflectance measurements to the surface normal directly, where inverse reflectance functions are approximated by a variety of designs of neural networks that are trained using a large set of synthesized images depicting arbitrary object appearances under various lighting conditions [22, 32].

Synthesizing images to train an inverse model from reflectance to surface geometry is usually restrictive. This is because to train a unified model that simultaneously quantifies material, shape and lighting, explicit reflectance modeling, such as inter-pixel constraints [8, 49], isotropy [22], parameterizations of rendering equations [46],etc., must be incorporated through human intervention. However, as the essence of photometric stereo is to resolve arbitrary surface shape using natural appearances, how to design an automatic, generic and per-pixel machinery that directly learns from reflectance measurements while without domain knowledge about material properties still merits further investigation.

To this end, our solution takes a different path. Rather than training a model in a completely supervised manner, we choose to train a simple network as a part of self-supervised learning pipeline in which Lambertian reflectance is the supervised signal. In fact, focusing on Lambertian reflectance for data-driven shape-material analysis is well-motivated: in theory, the intrinsic scene characteristics [3] are believed to form an essential component for shape perception; numerically, Lambertian signals encode surface normal almost linearly so flexible estimation is possible; representation-wise, since general reflectance and Lambertian reflectance co-reside in the same space, elimination of material properties can regarded as a transformation from the former to the latter. Explicitly parameterizing general reflectance as a subjective, perception-based measure without domain knowledge is difficult, but solving a simpler task that learns how general reflectance can be reduced to its Lambertian counterpart has the same effect.

Therefore, our objective is to (1) identify the subspace that only contains the signals of Lambertian reflectance, and (2) design a procedure that transforms the signals of general reflectance to this subspace. In other words, we demand a per-pixel mapping from the general appearances of a surface point to its Lambertian counterpart, and we introduce the concept of appearance flow to model this process. In particular, we want the flow to perform a many-to-one mapping and in our design it is implemented by a composition of mappings of two types: (1) cross-space mapping and (2) space-invariant mapping. Correspondingly, a transfer operator is introduced to implement an encode-transfer-decode procedure that allows the flow to contract in a latent space formed by the Lambertian reflectance. During training, each sample is associated with its corresponding Lambertian signal as a template, which is used to regularize the distribution of the latent signals. This training process is self-supervised as in its pretext task shape intrinsics are obtained without explicitly labeled material properties, and in its downstream task surface shape is recovered using Lambertian reflectance only. In our design there are two types of template signals: one used to regularize lighting distribution while the other represents the corresponding Lambertian reflectance under the designated lighting. As a result, the learned flow converts a sample under arbitrary lighting into the Lambertian appearances sampled under lights with fixed positions. An overview of our design is illustrated in Figure 1.

To sum up, in this paper, we introduce a generic semi-supervised learning pipeline that learns surface reflectance automatically from the tabulated optical measurements for photometric stereo. The contributions of this work include:

1. A theoretical framework to establish the concept of appearance flow that is motivated by the fundamental understanding about intrinsic scene characteristics in the context of shape perception.

2. A generic encode-transfer-decode procedure to implement flows in a systematic and consistent design.

3. A self-supervised training strategy that effectively ensures the flows to be contractive .

4. An effective and lightweight neural network for photometric stereo that is trained without domain-specific knowledge.

The remaining of this paper is organized as follows: Section 2 overviews the related work; Section 3 presents the formal definition of contractive appearance flow and the generic design of encode-transfer-decode procedure; Section 4 presents an effective training strategy for

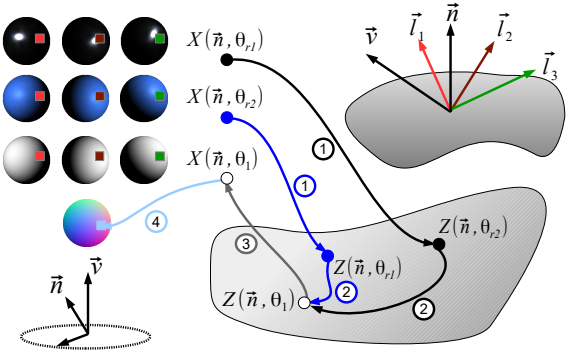

Figure 1: An overview of our design of per-pixel contractive flow for photometric stereo. The appearances of each pixel is represented as a vector of intensities (*e.g.* $X \in R^3$ in the case of three directional lights). In the above example, the flow defines a mapping from appearances of two materials (aluminium and blue rubber [36]), denoted by $X(\vec{n},\theta_{r_1})$ and $X(\vec{n},\theta_{r_2})$, to the latent signal $Z(\vec{n},\theta_1)$, which can be decoded to synthesize a set of Lambertian appearances, $X(\vec{n},\theta_1)$. The flow takes three steps: (1) encode, (2) transfer, (3) decode. The latent space is identified by the encoder, and the flow is **contractive** if the transfer operator allows all latent signals encoding the same surface geometry to converge. For instance, $Z(\vec{n},\theta_{r1}) \rightarrow Z(\vec{n},\theta_1)$, $Z(\vec{n},\theta_{r2}) \rightarrow Z(\vec{n},\theta_1)$, and $Z(\vec{n},\theta_1)$ is the point of contraction. Normal estimation is performed independently in a separate downstream task.

photometric stereo. Section 5 discusses the performance of our solution with experiment results. Section 6 concludes this paper.

2 Related Work

Photometric stereo [47] is a derived problem of shape from shading [20] in which inter-pixel constraints can be removed. So the reflectance model is the core of its solution [44]. Various models have been proposed but in recent years deep learning-based approaches become dominant.

2.1 Traditional Formulation

In the simplest case object appearances are confined in a low dimensional linear space by the Lambertian reflectance [6], which can be generalized by taking the specular signal as an outlier [48]. Various nonlinear models have also been proposed, where parameters are set algebraically [23, 42] or derived from physical or geometric models [11, 12, 16]. Their performance can be further enhanced by confining the parameters using examplers, which is done so by optimizations [18, 21] or through direct matching [15]. While some work addresses scenarios of near field lighting [31], the majority of cases assume directional light. The effect of lighting can be analyzed using bases of spherical harmonics [4, 39], and when the light is uncalibrated, low rank constraints [5, 38] is used to resolve ambiguity [7] to some extent. Derive isocontours as a geometric primitive from isotropic reflectance serve as useful

shape cue [2, 45], and algebraic formulation is applied in a less constrained environment [33, 34, 41], which assumes that object appearances and surface geometry lie on two correlated low-dimensional manifolds. Additionally, smoothness constraints may also be leveraged using variational approaches [17, 37].

2.2 Deep Learning-based Formulation

Neural networks are powerful in that they can represent a forward or inverse reflectance model flexibly. Due to their differentiability, they can be used to as a renderer to determine the surface normal by minimizing a their output with the actual observation [25, 46]; alternatively, they may also set up an inverse mapping from pixel values to surface normal directly [40]. To achieve flexible per-surface-point training, the concept of "observation map" is introduced [22] as a canonical 2D representation of the incident light field, between which and the surface normal an inverse mapping is established. On the other hand, Siamese network is an alternative to indirectly regularizing the image-based learning of reflectance functions [8]. In this case, image features inferring surface geometry are obtained through pooling. Furthermore, spline interpolation [50] and graph-based model are [49] available to address the case when distribution of the light is sparse. A neural network trained dedicated for calibration can be used in a cascade pipeline [9] and the concerns about ambiguity is discussed in [10]. Part of **our design** exhibits certain analogies to "observation map" [22] in that both work model per-pixel incident light field. The difference is that the input into our pipeline is a vector and its underlying implementation is all-MLP.

2.3 VAE and Representation Disentanglement

The theoretical essence of contractive appearance flow lies in structuring the distribution of latent signals encoded by a VAE [27] through constraints [19]. Since the latent space may represent a union of multiple subspaces with respect to some semantic meaning, and the ability to "disentangle" them is highly desirable [1]. This can be achieved through imposing low-rank constraint [14], explicitly modeling [24], learning subspaces in ensemble [35], or applying matrix factorization [30]. Labels may be utilized in various ways to facilitate the factorization process [13, 28, 29]. In this context of subspace decomposition,**our design** only focuses on one subspace, which is the manifold that only contains encoded Lambertian signals, $Z(\vec{n},\theta_1)$. Moreover, instead of identifying the structures of the supplement space, we train a transfer operator that associates the signals lie outside the subspace with their counterparts lying inside it. Moreover, contractive flow is non-invertible, which differentiates it from regular normalizing flows [26].

3 Contractive Appearance Flow

A reflectance function is defined in terms of three quantities: the surface normal $\vec{n}$, the directional light $\vec{l}$ and the direction of the view $\vec{v}$, and it is normally expressed in the form of

$$I_{\vec{l}} = f_r(\vec{l}^\intercal\vec{n}, \vec{l}^\intercal\vec{v}, \vec{n}), \tag{1}$$

where scalar $I_{\vec{l}}$ is the pixel-wise appearance which is in turn parameterized as $I_{\vec{l}}(\vec{n},\theta_r)$, where θ_r denotes a set of unknown parameters modeling a material's reflectance property. We use θ_1 to denote Lambertian reflectance: $I_{\vec{l}}(\vec{n},\theta_1) = \max(\vec{l}^\intercal\vec{n}, 0)$.

3.1 Formulation

Illuminating a surface point by a set of K directional lights $L = \{\vec{l}_1, \vec{l}_2, \ldots, \vec{l}_K\}$ produces a sequence of appearances represented by a K-tuple: $X(\vec{n}, \theta_r, L) = \{I_1, I_2, \ldots I_K\}_{\vec{n},\theta_r}$. As $\vec{n}$ and θ_r vary, a field of vectors is formed for any specific L. Unless otherwise stated, in the following literature L is omitted and treated as a constant. Over this field, we define a per-pixel flow:

$$F_{\vec{n}}(\theta_r, \theta_1) : X(\vec{n}, \theta_r) \rightarrow X(\vec{n}, \theta_1), \tag{2}$$

with $X(\vec{n}, \theta_1) = \{\vec{l}_1^{\intercal}\vec{n}, \vec{l}_2^{\intercal}\vec{n}, \ldots, \vec{l}_K^{\intercal}\vec{n}\}$. Moreover, $F_{\vec{n}}(\theta_r, \theta_1)$ is said to be **contractive** if it is guaranteed to reach its **point of contraction**, $X(\vec{n}, \theta_1)$, from all possible starting positions designated by θ_r.

Accordingly, we can define a contractive mapping, $\Phi(\cdot)$, over the field of X such that:

$$|\Phi(X(\vec{n}, \theta_r)) - \Phi(X(\vec{n}, \theta_1))| \leq |X(\vec{n}, \theta_r) - X(\vec{n}, \theta_1)| \tag{3}$$

where $|\cdot|$ is a type of norm designated by the loss function used for training (Section 4.2). It can be readily observe that $F_{\vec{n}}$ can be represented by a composition of contractive mappings: $F_{\vec{n}}(\theta_r, \theta_1) = \Phi_D \circ \Phi_{D-1} \ldots \circ \Phi_1(\cdot)$, with $X(\vec{n}, \theta_r)$ being the input to Φ_1 and $X(\vec{n}, \theta_1)$ being the output of Φ_D. It is worth noting that $\Phi(\cdot)$ does not depend on $\vec{n}$. In this case, we have two types of mapping: (1) cross-space mapping, which is implemented by an encoder and decoder that can alter the dimension of a signal representation; (2) space-invariant mapping, which alters the relative positions among signals in the same space. A standard VAE does the former, whereas we introduce the design of transfer operator to achieve the latter. Essentially, a contractive flow can be simply carried out by a combination of encoding, transfer and decoding procedure, all of which can be implemented by Multi-Layer Perceptrons (MLPs) along a single pipeline. In other words, we can assemble and train an all-MLP neural network to implement a contractive flow $F_{\vec{n}}$.

3.2 Encode-Transfer-Decode Procedure

Equation 1 indicates that for any X there exists a low-dimensional latent embedding parameterized by $\vec{n}$ and θ_r. This allows us to design an encoder, $\Phi_{en}(\cdot)$, to encode $X(\vec{n}, \theta_r)$ in a latent space containing $Z(\vec{n}, \theta_r)$:

$$\Phi_{en}(X(\vec{n}, \theta_r)) = Z(\vec{n}, \theta_r), \tag{4}$$

where $Z(\vec{n}, \theta_r)$ has a much lower-dimensional representation than $X(\vec{n}, \theta_r)$. We prefer to operate on $Z(\vec{n}, \theta_r)$ instead of $X(\vec{n}, \theta_r)$ because mappings between low dimensional signals can be implemented by simpler networks, which are easier to train.

However, direct encoding does not guarantee Equation 3 to hold. In fact, it is as challenging to obtain a VAE-based pipeline that achieves $Z(\vec{n}, \theta_r) = Z(\vec{n}, \theta_1)$ as to set up a direct inverse mapping discussed in Section 1. This is because in a directly-encoded space the variance of θ_r caused by material variations persist. Instead, appearance flow allows us to set up a space-invariant transfer mapping, $\Phi_{tr}(\cdot)$,

$$\Phi_{tr}(Z(\vec{n}, \theta_r)) = Z(\vec{n}, \theta_r'), \tag{5}$$

that eliminates these variations by enforcing the following:

$$|Z(\vec{n}, \theta_r') - Z(\vec{n}, \theta_1)| \leq |Z(\vec{n}, \theta_r) - Z(\vec{n}, \theta_1)|. \tag{6}$$

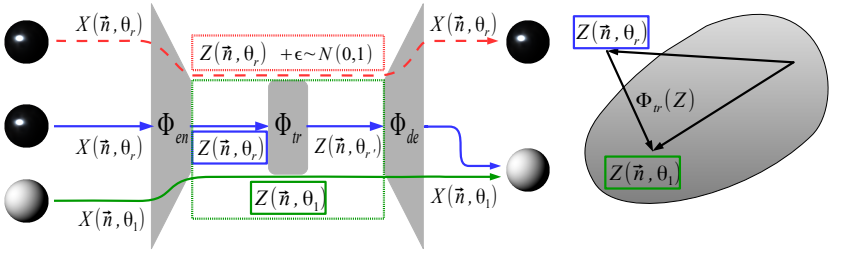

Figure 2: We design an encode-transfer-decode procedure to implement contractive flow. The dashed line represents a typical VAE pipeline for comparison. The difference is that instead of enforcing a multivariate gaussian distribution in the latent space, we introduce a transfer operator,$\Phi_{tr}(\cdot)$, that regularizes the distribution of the latent signals (in the middle) using an explicit sample $X(\vec{n},\theta_1)$ from the corresponding Lambertian appearances (at the bottom). The transfer operator essentially performs a nonlinear projection in the latent space through which signals are mapped to a subspace containing only $Z(\vec{n},\theta_1)$. We use a residual block to implement this process, where by subtraction we want $Z(\vec{n},\theta_r')$ and $Z(\vec{n},\theta_1)$ to converge. The residual signal being taken away encodes the information about material properties.

which regularizes the distribution of the encoded signals in place of the KL-divergence in the case of standard VAE. It is worth noting that this procedure is purely data-driven and no explicit modeling about θ is involved.

In particular, the operation described by Equation 5 and 6 exhibits certain analogy to a projection operator. Namely, the transfer operator decomposes $Z(\vec{n},\theta_r)$ into two subcomponents: one lies in the subspace spanned by $Z(\vec{n},\theta_1)$ with respect to $\vec{n}$, and one is "orthogonal to/independent of it". Our design creates a similar scheme by deploying a residual block denoted as $\Phi_{tr}(\cdot)$, which represents its input as a superposition of a signal of our interest and its residual. In a nonlinear context, we want the residual component to carry away the information uncertainty about the material properties (θ), and only the information about the Lambertian reflectance to remain. This intuition and analogy are illustrated in Figure 2. It is worth noting that each transfer operator is paired with an encoder, as the latter creates the latent space for the former to operate.

Furthermore, at the end of the pipeline a decoder, $\Phi_{de}(\cdot)$:

$$\Phi_{de}(Z(\vec{n},\theta_1)) = X(\vec{n},\theta_1), \tag{7}$$

is attached for two reasons. First, it prevents the distribution of the signal from collapsing in the latent space (*i.e.* creating a vanishing point) ; second, $\vec{n}$ can be obtained from $X(\vec{n},\theta_1)$ directly by solving a linear least sqaures problem in the downstream task. The subspace of $Z(\vec{n},\theta_1)$ is rank-3.

4 Training to Obtain Shape Intrinsics

To train a network that implements contractive appearance flows, a variety of combinations of L, $\vec{n}$ and θ_r should be included in the training data, with θ_r being an unobserved quantity.

Training mainly consists of two tasks: (1) design a robust and neural network-compatible representation to accommodate inputs of variable lengths; (2) design a training strategy that ensures the resulting flow is contractive. Even though the first task is not the main focus of this work, with the tools described in Section 3.2, we show that both tasks can be accomplished by a single learning pipeline end-to-end. Also, the training is self-supervised because among all the training samples only the ones sampled from Lambertian reflectance are identified (*i.e.* $\bar{X}(\vec{n},\theta_l)$ defined in Section 4.1) and serve as the supervisory signals.

The pipeline identifies two latent spaces, $\mathbf{Z}_1$ and $\mathbf{Z}_2$, where in $\mathbf{Z}_1$ through transfer we want per-pixel reflectance signals of the same material sampled by lights of varying distribution to have a consistent representation, and in $\mathbf{Z}_2$ we want all signals carried by the same per-pixel flow to meet at $Z_2(\vec{n},\theta_l)$. In other words, in each latent space we specify a type of point of contraction, $\bar{Z}_1(\vec{x},\theta_r)$ (Section 4.1) and $\bar{Z}_1(\vec{x},\theta_l)$ (Section 4.2), respectively, and our training objective is to enforce the flow to pass through them. To this end, we associate each sample with two types of template samples, $\bar{X}(\vec{x},\theta_r)$ and $\bar{X}(\vec{x},\theta_l)$, which are to be encoded into the two points of contraction,respectively. We make use of the fact that a transfer operator behaves as an **identity mapping** when its input signal is the point of contraction (*e.g.* it is the template signal) to instruct the transfer operators how to regularize the two latent subspaces. Specifically, we enforce the pipeline to minimize the **contractive loss**,L_1 and L_2, respectively. They have a similar effect to that of KL loss in the case of VAE. The performance of the decoder is measured by the **reconstruction loss** L_3. The training workflow is depicted in Figure 3.

4.1 A Consistent Representation for Incident Light Fields

In Section 3.1, we use $X(\vec{n},\theta_r)$ to denote a vector of pixel values whose length depends on L, but since in practise the distribution of L is not known as a priori, it needs a consistent representation in order to be neural network-compatible. Let $\bar{L}$ denote **a fixed set** of samplers over the spherical surface with which **standard representation**,$\bar{X}(\vec{n},\theta_r)$, is defined, and $\bar{Z}(\vec{n},\theta_r)$ be its latent signal. So, by enforcing $Z_1(\vec{n},\theta_r)\rightarrow\bar{Z}_1(\vec{n},\theta_r)$, the effect of varying L is reduced.

Solutions exist to address this issue [49, 50], but since both implementations are independent of the design of appearance flow, incorporating them into our design will be our future work. Instead, similar to the idea of "observation map" [22], we partition the lighting hemisphere into a set of disjoint blocks to generate $X(\vec{n},\theta_r)$ out of an arbitrary L. Accordingly, a transfer operator is trained by minimizing the function of contractive loss,

$$L_1 = |Z_1(\vec{n},\theta_r) - \bar{Z}_1(\vec{n},\theta_r)|, \tag{8}$$

and when the optimum is achieved, it becomes an identity mapping for $\bar{Z}_1(\vec{n},\theta_r)$.

4.2 Detect the Subspace of Lambertian Reflectance

Subsequently, we also want to detect the subspace only spanned by $\bar{Z}_2(\vec{n},\theta_l)$ in $\mathbf{Z}_2$. To this end, we include an additional template sample, $\bar{X}(\vec{n},\theta_l)$, whose encoded signal, $\bar{Z}_2(\vec{n},\theta_l)$ undergoes an identity mapping in $\mathbf{Z}_2$. This leads to the second contractive loss,

$$L_2 = |\frac{1}{2}Z(\vec{n},\theta_r') + \frac{1}{2}\bar{Z}(\vec{n},\theta_r') - \bar{Z}(\vec{n},\theta_l)| \tag{9}$$

which reflects the requirement that the flow should end up at the point of contraction $Z(\vec{n},\theta_l)$.

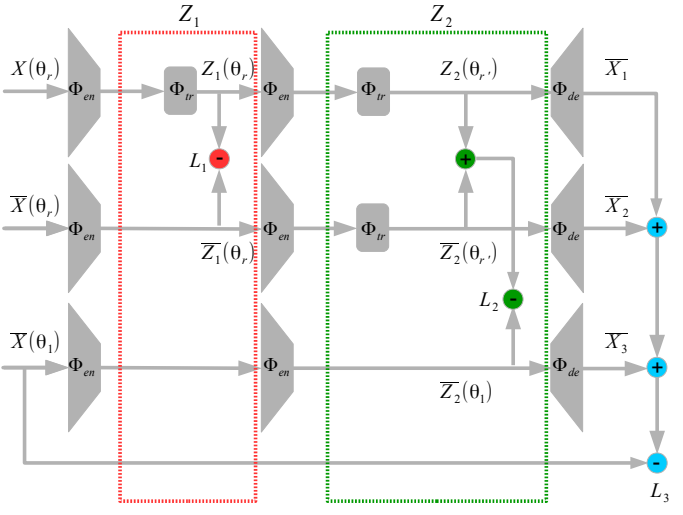

Figure 3: The workflow of the per-pixel self-supervised training process. $\vec{n}$ is omitted in notations for brevity. There are three data paths passing through two latent spaces Z_1 and Z_2, among which the top path: $X(\theta_r) \to Z_1(\theta_r) \to Z_2(\theta_r') \to \bar{X}_1$ represents the standard procedure used in testing. There are three samples, two of which are templates: $X(\theta_r)$ is arbitrarily sampled under L and standardized by the nearest neighbor matching against $\bar{L}$; $\bar{X}(\theta_r)$, the appearances with same reflectance properties sampled from $\bar{L}$, and $\bar{X}(\theta_1)$, the Lambertian appearances sampled also from $\bar{L}$. Correspondingly, L_1 and L_2 are contractive loss and L_3 is the reconstruction loss in terms of $X(\theta_1)$. $\bar{Z}(\vec{n},\theta_1)$ is the point of contraction.

Finally, we introduce a reconstruction loss,

$$L_3 = |\frac{1}{3}\bar{X}_1 + \frac{1}{3}\bar{X}_2 + \frac{1}{3}\bar{X}_3 - \bar{X}(\vec{n},\theta_1)|, \tag{10}$$

which produces $\bar{X}(\vec{n},\theta_r) = \max(\bar{L}^\intercal\vec{n},0)$, from which $\vec{n}$ is solved by a linear least-squares problem. The loss function is taken as the sum of L_1, L_2 and L_3 and they are all measured in L_2-norm. Because the decoding process is much simpler than the encode-transfer procedure, the architecture of the pipeline is asymmetric.

5 Experiment

Table 1 compares the performance of our solution (AF) with other recently-proposed per-pixel solutions to deep-learning-based photometric stereo that represent the state-of-the art. "AF-VAE" is the results of an ablation study. The experiment settings, the architecture of the network and performance analysis with illustrations can be found in the supplemental material.

	BALL	CAT	POT1	BEAR	POT2	BUDDHA	GOBLET	READING	COW	HARVEST	AVG.
CNN-PS[23]	2.2	4.6	5.4	4.1	6.0	7.9	7.3	12.6	8.0	14.0	7.2
DPSN[40]	3.4	7.2	7.9	7.2	8.8	13.3	12.3	17.4	8.4	16.8	10.3
BS	4.1	8.4	8.8	8.3	14.6	14.9	18.5	19.8	25.6	30.6	15.4
AF	3.3	7.8	7.8	7.5	9.8	13.4	12.4	15.9	9.8	18.7	10.6
AF-VAE	30.1	36.1	36.2	34.2	36.7	38.1	36.0	32.4	40.5	42.7	36.3

Table 1: Comparison of the benchmark results [43] produced by our solution and other per-pixel methods with baseline performance. Ablation study(AF-VAE): the entire pipeline is trained as if it is a standard VAE to generate Lambertian reflectance, where template signals are disabled in training. It can be seen that the transfer operator trained using the template signals plays a critical role in ensuring the flow to contract in the space spanned by Lambertian signals.

Figure 4: Our estimation results on DiLiGent image set [43]. From top to bottom: Pot1, Buddha, Reading and Harvest. The error is normalized and saturates at 40 degrees. It can be observed that cast shadows and interreflections are the main sources of estimation error. The reflectance displayed by various types of materials, be it diffusive or specular, is properly addressed.

5.1 Performance

Though the performance of our solution is inferior to CNN-PS, it is worth noting that in terms of model complexity, the resources consumed for training and the way the network is trained, our solution are advantageous in that it adopts a self-supervised pipeline that only consumes a small fraction of training data required by other methods (2GB v.s. 10+GB), but it generalizes reasonably well. In particular, our model is trained on monochromatic MERL [36] that only tabulates direct optical measurements. Since our model does not require domain-specific knowledge, the process for collecting training data would be much easier than synthesizing images with a renderer, which is a major requirement for other methods.

As illustrated by Figure 4, the major source of error of our estimations is inter-reflections ("Reading" and "Harvest") and cast shadows ("Buddha"). This indicates that extra components needed to be deployed to address these optical phenomena in addition to the pipeline designed for shape-reflectance analysis. Since the architecture of our network is extremely light weight, and the function of each block can be explicitly designated, the proposed design has the flexibility to scale-up to address these issues. This defines a major task of our follow-up work.

5.2 Ablation Study

We conducted an ablation study and include the results in the last line of Table 1 (AF-VAE). It is to prove the necessity of applying template signals in training and deploying the transfer block in our design. Namely, if we remove the constraints imposed by the template samples during training, and resort to a typical implementation of VAE with its encoder taking $X(\theta_r)$ as input and its decoder being trained to produce $\bar{X}(\theta_l)$, with the loss functions in Equation 8 and 9 being disabled, the implementation will fail to generalize to the testing data completely, as the results suggest. After all, without the template samples regularizing the distribution of the signals in the latent space, the neural network will not be able to capture the underlying geometry-dependent structure solely from end-to-end instruction.

6 Conclusion

We establish the concept of contractive appearance flow to address photometric stereo with general reflectance. The main idea of the proposed design is to learn a procedure that transforms a set of per-pixel appearances caused by general surface reflectance to its Lambertian counterparts, which is aligned with the traditional treatment of intrinsic shape recovery from material reflectance but offers a generic data-driven tool set. In particular, we propose a self-supervised learning pipeline implemented with generic encode-transfer-decode procedure, and by placing the point of contraction in the geometry-dependent latent space we are able to synthesizes Lambertian appearances without any explicit modeling or parameterizations regarding the material properties that often involve domain-specific knowledge. Our solution is validated by benchmark experiments on real world data.

Though, our method does not deliver the state-of-the-art performance at the moment, this new framework offers plenty of room for improvement. By integrating a mechanism that allows for flexible representations for the incident light field in our follow-up work, we expect this design to be applied to a broader range of applications in which cast shadows and interreflections will be be addressed.

References

[1] Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, 19, 2018.

[2] Neil G Alldrin and David J Kriegman. Toward reconstructing surfaces with arbitrary isotropic reflectance: A stratified photometric stereo approach. In *Proc. of International Conference on Computer Vision*, 2007.

[3] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. Vis. Syst*, 2, 1978.

[4] Ronen Basri and David W Jacobs. Lambertian reflectance and linear subspaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(2), 2003.

[5] Ronen Basri, David Jacobs, and Ira Kemelmacher. Photometric stereo with general, unknown lighting. *International Journal of Computer Vision*, 72(3), 2007.

[6] Peter N Belhumeur and David J Kriegman. What is the set of images of an object under all possible illumination conditions? *International Journal of Computer Vision*, 28(3), 1998.

[7] Peter N Belhumeur, David J Kriegman, and Alan L Yuille. The bas-relief ambiguity. *International Journal of Computer Vision*, 35(1), 1999.

[8] Guanying Chen, Kai Han, and Kwan-Yee K Wong. PS-FCN: A flexible learning framework for photometric stereo. In *Proc. of European Conference on Computer Vision*, 2018.

[9] Guanying Chen, Kai Han, Boxin Shi, Yasuyuki Matsushita, and Kwan-Yee K Wong. Self-calibrating deep photometric stereo networks. In *Proc. of Computer Vision and Pattern Recognition*, 2019.

[10] Guanying Chen, Michael Waechter, Boxin Shi, Kwan-Yee K Wong, and Yasuyuki Matsushita. What is learned in deep uncalibrated photometric stereo? In *European Conference on Computer Vision*, 2020.

[11] Lixiong Chen, Yinqiang Zheng, Boxin Shi, Art Subpa-asa, and Imari Sato. A microfacet-based model for photometric stereo with general isotropic reflectance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[12] Hin-Shun Chung and Jiaya Jia. Efficient photometric stereo on glossy surfaces with wide specular lobes. In *Proc. of Computer Vision and Pattern Recognition*, 2008.

[13] Antonia Creswell, Yumnah Mohamied, Biswa Sengupta, and Anil A Bharath. Adversarial information factorization. *arXiv preprint arXiv:1711.05175*, 2017.

[14] Zhengming Ding and Yun Fu. Deep domain generalization with structured low-rank constraint. *IEEE Transactions on Image Processing*, 2017.

[15] Kenji Enomoto, Michael Waechter, Kiriakos N. Kutulakos, and Yasuyuki Matsushita. Photometric stereo via discrete hypothesis-and-test search. In *Proc. of Computer Vision and Pattern Recognition*, 2020.

[16] Athinodoros S Georghiades. Incorporating the Torrance and Sparrow model of reflectance in uncalibrated photometric stereo. In *Proc. of Computer Vision and Pattern Recognition*, 2003.

[17] Bjoern Haefner, Zhenzhang Ye, Maolin Gao, Tao Wu, Yvain Quéau, and Daniel Cremers. Variational uncalibrated photometric stereo under general lighting. In *Proc. of International Conference on Computer Vision*, 2019.

[18] Aaron Hertzmann and Steven M Seitz. Example-based photometric stereo: Shape reconstruction with general, varying brdfs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8), 2005.

[19] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. 2016.

[20] Berthold KP Horn. Shape from shading: A method for obtaining the shape of a smooth opaque object from one view. 1970.

[21] Zhuo Hui and Aswin C Sankaranarayanan. Shape and spatially-varying reflectance estimation from virtual exemplars. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(10), 2017.

[22] Satoshi Ikehata. Cnn-ps: Cnn-based photometric stereo for general non-convex surfaces. In *Proc. of European Conference on Computer Vision*, 2018.

[23] Satoshi Ikehata, David Wipf, Yasuyuki Matsushita, and Kiyoharu Aizawa. Robust photometric stereo using sparse regression. In *Proc. of Computer Vision and Pattern Recognition*, 2012.

[24] Maximilian Ilse, Jakub M Tomczak, Christos Louizos, and Max Welling. Diva: Domain invariant variational autoencoders. In *Medical Imaging with Deep Learning*, 2020.

[25] Berk Kaya, Suryansh Kumar, Carlos Oliveira, Vittorio Ferrari, and Luc Van Gool. Uncalibrated neural inverse rendering for photometric stereo of general surfaces. *arXiv preprint arXiv:2012.06777*, 2020.

[26] Diederik P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in Neural Information Processing Systems*, 2018.

[27] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *Proc. of International Conference on Learning Representations*, 2014.

[28] Diederik P Kingma, Danilo J Rezende, Shakir Mohamed, and Max Welling. Semi-supervised learning with deep generative models. *Advances in Neural Information Processing Systems*, 2014.

[29] Jack Klys, Jake Snell, and Richard Zemel. Learning latent subspaces in variational autoencoders. *Advances in Neural Information Processing Systems*, 2018.

[30] Xiao Li, Chenghua Lin, Ruizhe Li, Chaozheng Wang, and Frank Guerin. Latent space factorisation and manipulation via matrix subspace projection. In *Proc. of International Conference on Machine Learning*, 2020.

[31] Fotios Logothetis, Roberto Mecca, Yvain Quéau, and Roberto Cipolla. Near-field photometric stereo in ambient light. 2016.

[32] Fotios Logothetis, Ignas Budvytis, Roberto Mecca, and Roberto Cipolla. Px-net: Simple and efficient pixel-wise training of photometric stereo networks. *arXiv preprint arXiv:2008.04933*, 2020.

[33] Feng Lu, Yasuyuki Matsushita, Imari Sato, Takahiro Okabe, and Yoichi Sato. From intensity profile to surface normal: photometric stereo for unknown light sources and isotropic reflectances. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10), 2015.

[34] Feng Lu, Xiaowu Chen, Imari Sato, and Yoichi Sato. Symps: Brdf symmetry guided photometric stereo for shape and light source estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1), 2017.

[35] Massimiliano Mancini, Samuel Rota Bulò, Barbara Caputo, and Elisa Ricci. Best sources forward: domain generalization through source-specific nets. In *IEEE international conference on image processing*, 2018.

[36] Wojciech Matusik and Matt Brand. A data-driven reflectance model. *ACM Transactions on Graphics (Proc. of ACM SIGGRAPH)*, 22(3), 2003.

[37] Yvain Quéau, Tao Wu, François Lauze, Jean-Denis Durou, and Daniel Cremers. A non-convex variational approach to photometric stereo under inaccurate lighting. In *Proc. of Computer Vision and Pattern Recognition*, 2017.

[38] Ravi Ramamoorthi. Analytic PCA construction for theoretical analysis of lighting variability in images of a Lambertian object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(10), 2002.

[39] Ravi Ramamoorthi and Pat Hanrahan. A signal-processing framework for inverse rendering. In *Proceedings of the 28th annual conference on Computer graphics and interactive techniques*, 2001.

[40] Hiroaki Santo, Masaki Samejima, Yusuke Sugano, Boxin Shi, and Yasuyuki Matsushita. Deep photometric stereo networks for determining surface normal and reflectances. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[41] Imari Sato, Takahiro Okabe, Qiong Yu, and Yoichi Sato. Shape reconstruction based on similarity in radiance changes under varying illumination. In *Proc. of International Conference on Computer Vision*, 2007.

[42] Boxin Shi, Ping Tan, Yasuyuki Matsushita, and Katsushi Ikeuchi. Bi-polynomial modeling of low-frequency reflectances. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(6), 2014.

[43] Boxin Shi, Zhipeng Mo, Zhe Wu, Dinglong Duan, Sai-Kit Yeung Yeung, and Ping Tan. A benchmark dataset and evaluation for non-Lambertian and uncalibrated photometric stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 2019.

[44] William M Silver. *Determining shape and reflectance using multiple images*. PhD thesis, Massachusetts Institute of Technology, 1980.

[45] Ping Tan, Satya P Mallick, Long Quan, David J Kriegman, and Todd Zickler. Isotropy, reciprocity and the generalized bas-relief ambiguity. In *Proc. of Computer Vision and Pattern Recognition*, 2007.

[46] Tatsunori Taniai and Takanori Maehara. Neural inverse rendering for general reflectance photometric stereo. In *Proc. of International Conference on Machine Learning*, 2018.

[47] Robert J Woodham. Photometric method for determining surface orientation from multiple images. *Optical engineering*, 19(1), 1980.

[48] Lun Wu, Arvind Ganesh, Boxin Shi, Yasuyuki Matsushita, Yongtian Wang, and Yi Ma. Robust photometric stereo via low-rank matrix completion and recovery. In *Proc. of Asian Conference on Computer Vision*, 2010.

[49] Zhuokun Yao, Kun Li, Ying Fu, Haofeng Hu, and Boxin Shi. Gps-net: Graph-based photometric stereo network. *Advances in Neural Information Processing Systems*, 33, 2020.

[50] Qian Zheng, Yiming Jia, Boxin Shi, Xudong Jiang, Ling-Yu Duan, and Alex C Kot. Spline-net: Sparse photometric stereo through lighting interpolation and normal estimation networks. In *Proc. of International Conference on Computer Vision*, 2019.

[51] Qian Zheng, Boxin Shi, and Gang Pan. Summary study of data-driven photometric stereo methods. *Virtual Reality & Intelligent Hardware*, 2020.