PhIT-Net: Photo-consistent Image Transform for Robust Illumination Invariant Matching

Damian Kaliroff
dkaliroff@technion.ac.il
Guy Gilboa
guy.gilboa@ee.technion.ac.il

Technion - Israel Institute of Technology
Haifa, Israel

Abstract

We propose a new and completely data-driven approach for generating a photo-consistent image transform. We show that simple classical algorithms which operate in the transform domain become extremely resilient to illumination changes. This considerably improves matching accuracy, outperforming the use of state-of-the-art invariant representations as well as new matching methods based on deep features. The transform is obtained by training a neural network with a specialized triplet loss, designed to emphasize actual scene changes while attenuating illumination changes. The transform yields an illumination invariant representation, structured as an image map, which is highly flexible and can be easily used for various tasks.

1 Introduction

Image processing and computer vision (CV) tasks often benefit from representations which are invariant to certain image changes. Photo-consistency is a highly desired property, essential for tasks based on color and contrast cues, such as matching, registration and recognition. Traditionally, illumination invariant representations were designed in a model-based manner. Lately, with the rise of deep learning, new data-driven algorithms are proposed to solve the problem. However, the physical assumptions and the models used for the data-driven approaches appear to limit their performance. We thus seek a very general, unconstrained, learning approach.

In this paper we propose a new paradigm for generating an illumination invariant image map. It is an unconstrained representation, generated in a self-supervised manner, completely data-driven, without using limiting inaccurate assumptions, such as the Lambertian model. We impose mild scale-consistency and geometrical constraints. The surprising representation derived by the training process provides new insights on invariant representations for matching. It can be used as a pre-processing stage for a wide variety of classical and learning-based algorithms, making them considerably more robust to lighting conditions.

To accomplish this, we design a deep neural network, referred to as *PhIT-Net* (**Ph**oto-consistent **I**mage **T**ransform **Net**work). It is trained in a self-supervised manner, using multiple sets of images of the same scene under different illuminations. This is illustrated in Fig. 1. We validate our proposed transform by various means. First, we show that images of the same scene, illuminated differently, are indeed represented in a very similar manner. This is

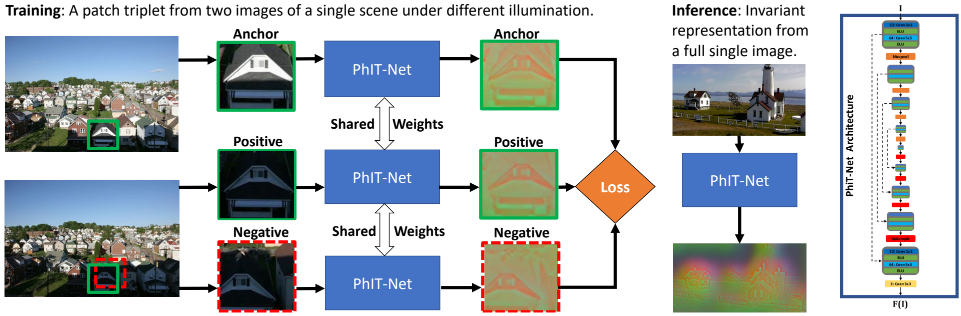

Training: A patch triplet from two images of a single scene under different illumination.
Anchor
Positive
Negative
PhIT-Net
Shared
Weights
Anchor
Positive
Negative
Loss
Inference: Invariant representation from a full single image.
PhIT-Net
PhIT-Net Architecture
F(I)

Figure 1: PhIT-Net learns a transformation using patch triplets taken from image pairs under different illumination conditions. In inference time the transformation creates an invariant representation from a single input image. The architecture is based on U-Net.

compared to other representations which seek illumination invariance. Next, we investigate the usability of our approach. Quantitative experiments are performed for patch matching and for rigid registration. Results are compared to state-of-the-art photo-consistent representations and to novel algorithms based on deep features. In both cases, we show our approach consistently yields superior results.

2 Related Work

There are two main branches of photo-consistent representations. The first approach attempts to estimate a physical quantity, the albedo (or reflectance) of objects in the image. Since the albedo is not affected by illumination and shading, it is inherently photo-consistent. A second branch is based on photo-consistent transforms, which serve to improve computer vision tasks, such as matching or registration. Our approach belongs to the latter.

Seeking the elusive albedo. Finding an intrinsic image representation is a long standing problem in computer vision. In [1, 20] the Retinex theory was introduced, followed by numerous algorithms, such as [8, 17, 18, 26, 28], with the aim of estimating reflectance and shading from a single image. Following the model by Barrow et al. [1], which assumes a Lambertian world, an image I is decomposed by $I = A \cdot S$, where A is albedo and S is shading. When this decomposition is based on a single image, it is referred to as SIID (Single Image Intrinsic Decomposition) [3, 21, 24]. Obtaining the albedo with SIID techniques is a hard ill-posed problem. Recent self-supervised deep learning algorithms attempt to learn this decomposition using extensive image data. In [21], Lettry et al. created a synthetic dataset of scenes with images under different illumination and trained a Siamese network [5] to decompose images into albedo and shading. In [22], Li and Snavely learn an albedo-shading decomposition by using natural photos in a dataset referred to as "BigTime" of indoor and outdoor scenes, each having several images with different lighting conditions. Recently, in [23] a dataset composed of Google street time-lapses is used to build an intrinsic decomposition approach. This method can work with time-lapses also at test time. They demonstrate their approach for the task of artificial scene relighting. Both [21] and [22] evaluate their results against ground truth intrinsic datasets, e.g. [3, 14]. The applicability of their albedo estimation for improving the performance of computer vision tasks is not tested. We use the BigTime dataset to develop our proposed photo-consistent transform.

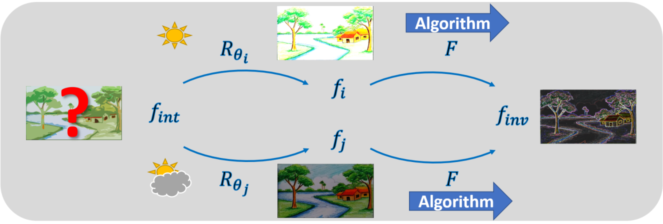


Figure 2: The invariant representation f_{inv} is generated by applying a transform F to an image f_k. Ideally, all images f_k, $k = i, j, ..$, emerging from the same intrinsic image f_{int}, by different transformations R_{θ_k} (image conditions), are mapped to the same invariant representation.

Seeking photo-consistency. Computer vision algorithms for matching, registration and optical-flow often assume a degree of object photo-consistency. In practice, however, illumination changes and shading yield an inconsistent representation in the raw color-space. This can be dealt with by either designing much more complex algorithms or by applying a pre-processing transform to the images, which is specifically targeted to increase photo-consistency. The latter approach has often shown to yield robust results, keeping the main CV algorithm simple and fast. We adopt this approach in our work.

In [15, 27] the census transform is applied to images, and serves as input for optical-flow computation, improving the robustness to illumination changes. In [13] and recently in [35], wavelet based pre-processing methods are used to remove the illumination component in face images to improve face recognition algorithms. For many recognition tasks, certain locations should be detected under different lighting conditions. Following the model of Finlayson et al. [10, 11] the authors of [25] and [31] propose single-channel illumination invariant representations of color images to improve place recognition, visual localization and classification algorithms.

We observed that strong illumination changes do not admit the albedo-shading assumptions of the SIID model. Our experiments indicate that albedo estimations are not very robust. We thus direct our efforts to finding an unconstrained photoconsisent map to be used by computer vision algorithms in varying lighting environments.

3 Unconstrained Invariant Representations

In this section we formalize the concept of an invariant representation which is unconstrained by a physical model. Let R_θ be some image transformation with a set of parameters θ. The transformation represents different conditions in which the image was acquired. It can model different attributes, such as illumination changes, noise, fog or atmospheric disturbances. Let f_{int} be an *intrinsic* image representation. An image instance f_i is obtained by applying the transformation R_{θ_i}, with specific parameters θ_i to the intrinsic image,

$$f_i = R_{\theta_i}(f_{int}). \tag{1}$$

For the intrinsic representation problem, the aim is to estimate f_{int}, given a single or multiple instances f_i in a blind manner, that is - without knowing θ_i. This is a difficult ill-posed problem. As an example, for the SIID problem, $I = f_i$, $f_{int} = A$, $\theta_i = S$ is some specific shading component and $R_\theta(f) = \theta \cdot f$.

In our approach, we aim only at obtaining an *invariant* representation. Thus, we do not seek to estimate θ_i and f_{int}. Instead, we would like to obtain a transform applied to an image instance, which is invariant to R_θ. This transform is denoted as $F(\cdot)$, and its purpose is depicted in Fig. 2. In the ideal case, all image instances f_i, created by R_{θ_i} applied to a specific intrinsic image f_{int}, are mapped by $F(\cdot)$ to the same invariant image, denoted as f_{inv}. Moreover, if two intrinsic scenes can be distinguished $d(f_{int}, g_{int}) > \varepsilon$, where $d(\cdot,\cdot)$ is some metric, so do the respective mappings $d(F(f_i), F(g_i)) > \delta$.

We now formulate the requirements for approximating this concept. Let $\mathbf{f} = \{f_1, .., f_N\}$ be a set of N instances of the same scene f_{int}, taken under different conditions, where f_i is defined by Eq. (1). Let $\mathbf{g}$ be defined in a similar manner with respect to a different scene g_{int}. Then F admits the following properties:

$$D(F(f_i), F(f_j)) \leq \varepsilon, \tag{2}$$

$\forall i, j = 1, ..., N$, where $D(\cdot,\cdot)$ is some distance and ε is a small constant. In addition,

$$D(F(f_i), F(g_j)) \geq c \cdot D(f_i, g_j), \tag{3}$$

$\forall i, j = 1, ..., N$, where $c \gg \varepsilon$ is some positive constant. A transform F admitting the above properties yields an unconstrained representation (not limited by formation models), approximately invariant under the transformation R_θ,

$$f_{inv,i} = F(f_i), \tag{4}$$

where $f_{inv,i} \approx f_{inv,j}$, $\forall i, j = 1, ..., N$. This transform minimizes the difference of images depicting the same scene (created from the same intrinsic image), and emphasizes differences between images from different scenes.

We introduce an additional requirement, which states that the properties above approximately hold for any part of the image. More formally, let us define a cropping operation of the image $crop_X$, where $X = (x_1, x_2, y_1, y_2)$ defines the cropping coordinates. We would like

$$crop_X(f_{inv,i}) \approx crop_X(f_{inv,j}),\ \forall i, j = 1, ..., N. \tag{5}$$

Moreover, to preserve the geometrical structure, it is desired that the crop operation also approximately commutes with F, that is

$$crop_X(F(f)) \approx F(crop_X(f)). \tag{6}$$

The combination of both requirements, Eqs. (5) and (6), can be written as,

$$crop_X(F(f_i)) \approx F(crop_X(f_j)),\ \forall i, j = 1, ..., N. \tag{7}$$

In order for Eq. (6) to be meaningful, f and $F(f)$ should have the same spatial dimensions. We refer to such a spatial representation as a *map*. For an input image of n pixels with k_i channels, the output is a map of n pixels with k_o channels, where k_o is a free parameter. We thus have $F : \mathbb{R}^{n \times k_i} \rightarrow \mathbb{R}^{n \times k_o}$. In order to obtain the transform F we do not need to directly model R_θ. We assume to have a training set comprised of M sets $\mathbf{f^m}$, $m = 1, ..., M$, each comprised of N instances of the same scene f^m_{int} transformed by $R_{\theta^m_i}$. We train a network that takes as input an instance f^m_i and produces an output $F(f^m_i)$, using a triplet network model [17, 19], following Eqs. (2) and (3). Additional losses are required to obtain a well-behaved, geometrically-consistent, sharp representation with several channels, as detailed above. This general methodology can be applied to develop representations invariant to different nuisance attributes. In this work, we develop a photo-consistent transform by applying this approach, obtaining an illumination invariant representation.

4 Application of Proposed Framework

We apply the framework presented in Section 3 for the problem of designing a photo-consistent transform. In this context, the transformations R_θ model different illumination conditions and our aim is to find $F(\cdot)$, such that it admits Eqs. (2)-(6). We use a neural network to compute F. We attempt to decrease the distance in the representation space between two corresponding patches of the same region in the scene, acquired at different lighting conditions. The training and inference procedures are illustrated schematically in Fig. 1. Our code is publicly available on GitHub https://github.com/dkaliroff/phitnet.

CNN Architecture. The CNN architecture of PhIT-Net is designed as an encoder-decoder U-net network [29]. The encoder and decoder are constructed using the same convolutional inception-like layers [33]. After the last decoder block, there is a 3x3 convolution layer to generate the final representation. The number of convolutions in this layer is determined by the number of channels in the final representation (three in our model).

Training Process. We train our network using a triplet network training scheme [17, 19]. In this scheme three instances of the same network are trained with shared weights. The input to the model is called a *patch triplet*. Each triplet is extracted from a pair of aligned images I_1, I_2, of the same scene, under different illumination conditions. For training, we use patches of size 64×64 pixels. The triplet is composed of an anchor patch (A), a positive patch (P) and a negative patch (N). The patches are defined as follows:

Anchor Patch (A). A random patch extracted from I_1, with a standard deviation above some threshold σ_p. We chose $\sigma_p = 25$ for the entire dataset (pixel values are in the range $[0, 255]$).

Positive Patch (P). A patch extracted from I_2, with the same coordinates as the Anchor, such that both patches depict the same scene (the RGB difference should stem mainly from illumination differences). In the representation space we would like (A) and (P) to be similar.

Negative Patch (N). A patch extracted from I_2 with shifted coordinates, relative to the positive patch. The shift is of 8 pixels, with respect to the anchor, with randomly chosen direction. The overlap induces a challenging and meaningful learning process.

Inference. At inference, a full input image is first passed through a single instance of the network, yielding floating point values in an arbitrary range. In order to reach an 8-bit image-format, as in the original images, we normalize the values over all channels linearly such that the minimum is mapped to 0 and the maximum to 255.

Training and Test Data. We use two different sets for outdoor and indoor settings. The main dataset is the outdoors dataset. It is composed of images from the BigTime dataset (See Fig. 3). We also train and evaluate our model on a set of indoor images. More details about the datasets and evaluation on the additional indoors dataset are provided in the supplementary material. The training is based on square patches of 64×64 pixels. The training set is composed of $240K$ triplets, extracted from 600 image pairs of 10 outdoor scenes. The evaluation was done using 100 image pairs selected from 17 additional outdoor scenes not used in training.

Loss Functions

The main loss function for the training process is the triplet loss [16, 36]. It aims at minimizing the distance between (A) and (P), while maximizing (up to a margin) the distance between (A) and (N). In order to reach a meaningful representation additional losses are

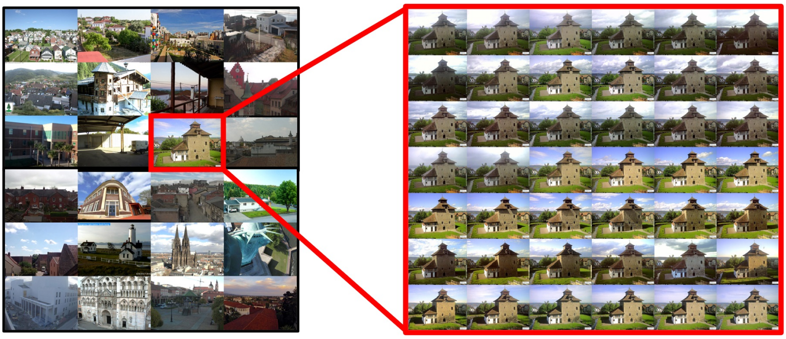


Figure 3: BigTime [22] dataset. Each scene is acquired under several lighting conditions.

required. These enable us to achieve some desired properties of the representation, such as scale consistency and channel variability. Let (f_a, f_p, f_n) be a triplet of image patches corresponding to (Anchor, Positive, Negative), respectively. Let $F(\cdot)$ be the output of PhIT-Net. Let $D_i(\cdot,\cdot)$ be a distance function. The total loss function is defined as a weighted sum of the following loss functions:

Triplet Loss (Inter-Loss):

$$L_T(f_a, f_p, f_n) = \max\{0, D_{corr}(F(f_a), F(f_p)) - D_{corr}(F(f_a), F(f_n)) + M\}, \tag{8}$$

where M is the triplet-loss margin (we use $M = 0.1$). Since patch affinity is often defined by correlation, we used the correlation distance function,

$$D_{corr}(x_1, x_2) = 1 - \frac{x_1 \cdot x_2}{\|x_1\|_2 \cdot \|x_2\|_2}. \tag{9}$$

Intra-Loss:

$$L_I(f_a, f_p) = D_{corr}(f_a, f_p) + \|f_a - f_p\|_2^2. \tag{10}$$

This loss promotes low A-P distance (in addition to the triplet loss), as suggested by [6]. Our experiments verify that adding this loss to the main triplet loss indeed improves performance. It also allows to minimize an additional distance function, not used in the main triplet loss.

Scale Consistency Loss:

$$L_{SC}(f_a) = D_{scale}(F(G(f_a, \rho)), G(F(f_a), \rho)), \tag{11}$$

where G is "Up-Sample and Crop" and represents a bilinear up-sampling by a random factor $\rho \in (1, 2]$ followed by a crop to the size of the original patch. The goal of this function, following Eq. (6), is to make the representation close to commutative with respect to these operations, as real images are. We use $D_{scale} = D_{corr}$.

Multi-Channel (MC) Similarity Loss: Let $I = F(f)$ be a multi-channel representation of K channels, $I = (I_1, ..I_K)$. The multi-channel loss is,

$$L_{MC}(I) = \sum_i \sum_{j \neq i} (1 - D_{corr}(I_i, I_j))^2. \tag{12}$$

We want the multi-channel representation to have significant and different information in each channel. Thus, we penalize channel similarity.

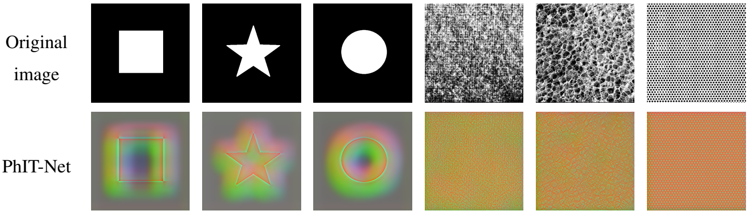


Figure 4: Basic shapes and Brodatz [32] textures and their corresponding transform.

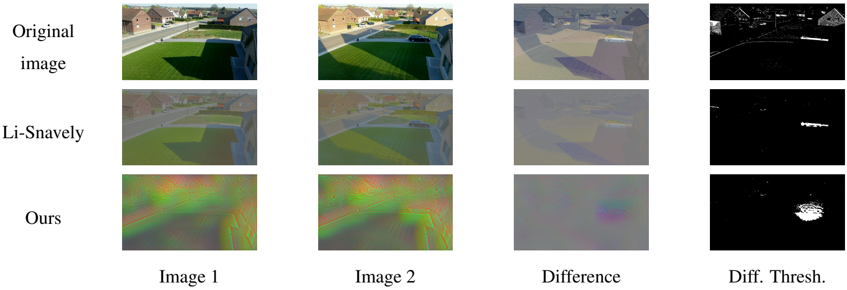


Figure 5: Actual scene differences (the car in this example) can be clearly detected under our transform, while illumination differences are attenuated.

5 Evaluation

We first examine the nature of the representation and its basic properties. As a sanity check, we verify that two images of the same scene under different illuminations are similar in the representation space. We then demonstrate the usability of our representation and show it improves the results of common computer vision tasks. Two tasks are examined quantitatively: patch matching and rigid registration.

Insights on the New Representation

Textures, Shapes and Color-Coding. In Fig. 4 we show the results of the proposed transform applied on basic shapes and textures. First, piecewise-constant shapes are examined. It is evident that edges are clearly defined. We observe that a certain color-coding is created. In flat regions the color provides information on the direction and distance of a nearby dominant edge, in a similar manner to the Chamfer distance transform [2]. We interpret this as means to disambiguate better flat regions with little variance. This property can assist matching algorithms. In addition, we examine textures with varying intensity. A highly uniform textural output is obtained in the representation space. Note that such textures do not appear in the training set. This demonstrates the generalization strength of PhIT-Net.

Visual Photo-consistency. In Fig. 5 we show that actual scene changes (the car) can be clearly seen, while illumination differences are attenuated. We compare this change detection experiment to the original images and to the representation of [22]. In Fig. 8 two examples of image pairs from the same scene under different illumination conditions are shown. All image channels in all representations are in the range $[0,255]$. Our representation has the lowest difference compared to other invariant representations (since the range is similar, no scaling is performed on the difference images). In order to demonstrate that

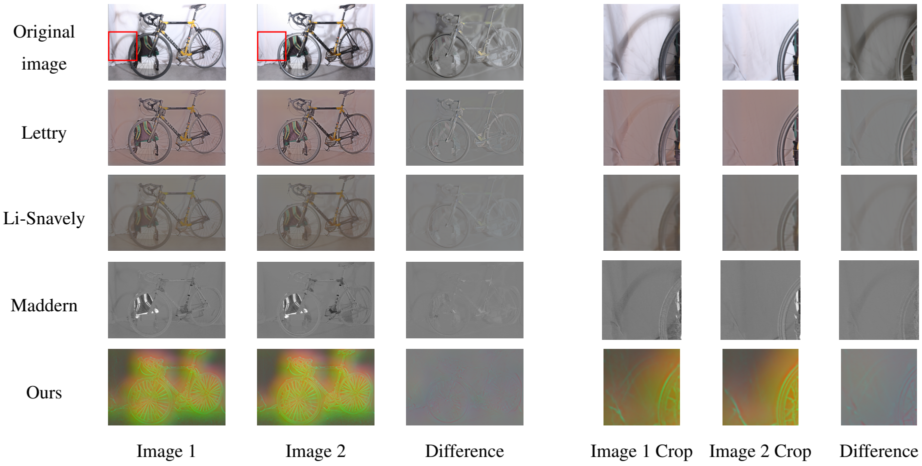


Figure 6: Visual comparison of invariant representation methods. A scene is shown under different illumination conditions. The difference (ideally zero) affirms that our representation is highly stable under illumination changes (zero is gray). Enlarged crops of the marked red-square are shown on the right.

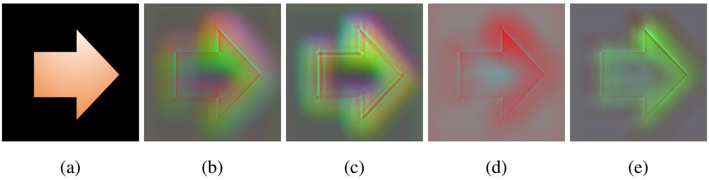


Figure 7: Ablation. (a) Original image, (b) Full model representation, (c) No Scale consistency loss, (d) No Multi-channel similarity loss, (e) With Rotation invariance loss.

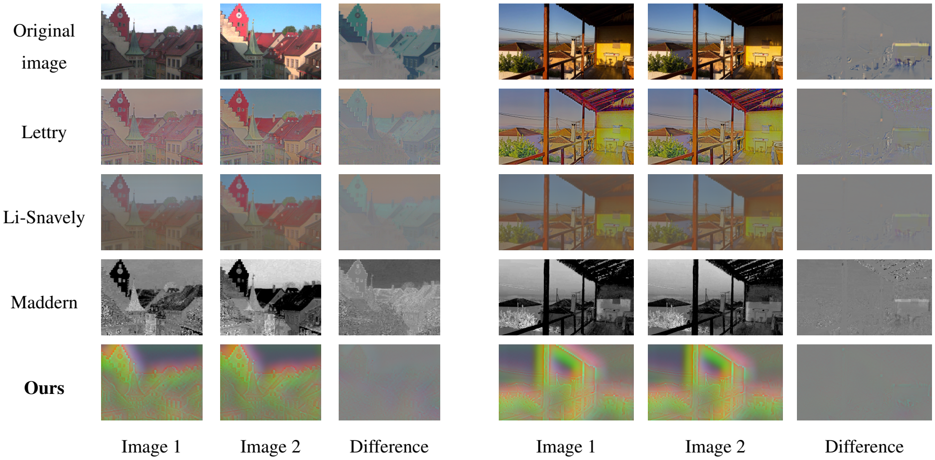


Figure 8: Visual comparison of invariant representation methods. For each scene, the representation of two images under different illumination conditions is shown. The difference (ideally zero) affirms that our representation is highly stable under illumination changes.

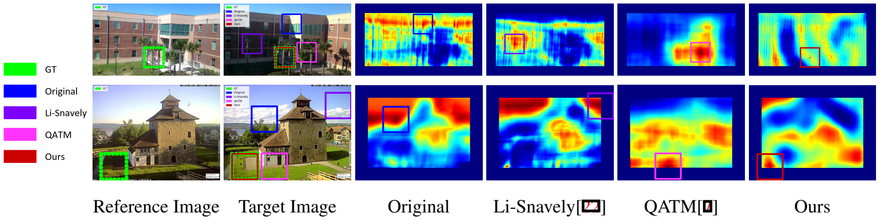


Figure 9: Patch matching of challenging scenes. Reference image on left (green frame). The frames marking the algorithms' results are overlaid on the target image. Heatmaps of each algorithm (right) indicate high (red) to low (blue) matching scores.

an unconstrained model can be beneficial compared to the standard intrinsic decomposition (albedo-shading) we show in Fig. 6 an example of an indoor scene from the Middlebury dataset [30]. It can be seen that our representation handles well strong shades on the wall as well as shiny metal, which violates the Lambertian model.

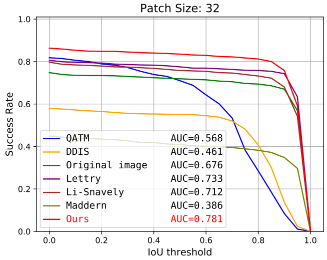
Ablation study. We performed an ablation study exploring different options of loss functions for the training of our network and also a representation with different channels in its output. We show in Fig. 7 an example of modifying the loss function and refer the readers to the supplementary material for a full ablation study with quantitative and qualitative results.

5.1 Quantitative Evaluation

We test our representation using two common computer vision tasks, and compare it with two unsupervised data-driven SIID methods, *Li-Snavely* [22] and *Lettry* [21], an analytic grayscale representation, *Maddern*, [25] and also with the *Original image*. In all cases the different representations are used as a pre-processing stage.

Patch Matching. In this task a template patch is selected from a reference image and the aim is to find its location in a target image. Both images are of the same scene but with different illumination conditions. We used the standard template matching function of OpenCV [4], *matchTemplate*, with the normalized cross correlation method. From each reference image, 10 square patches are randomly selected from significant areas in the image, by setting a minimum standard deviation of 25 (image range is $[0,255]$). This is done for three different patch sizes (32, 64 and 128 pixels). We compare our results to others invariant representation (using the same matching algorithm) and also to state-of-the-art dedicated template matching algorithms: QATM [7] and DDIS [34]. The latter are novel algorithms, based on deep features.

In Fig. 9 some matching results are shown, along with correlation-based heatmaps, which correspond to the closest match. Whereas the algorithms are generally robust to minor illumination changes, in these challenging cases, only our proposed transform succeeds. In Fig. 10 results of extensive quantitative experiments are shown. Accuracy is measured by intersection over union (IoU). Plots show the IoU-ROC curves and the area under the curve (AUC) scores for all algorithms. On the right of Fig. 10 a summary of the AUC score is given for all patch sizes. Our representation consistently achieves the highest score for all patch sizes. We conclude that using classical patch-matching algorithms jointly with our proposed transform (as pre-processing) surpasses not only other invariant transforms but also state-of-the-art end-to-end dedicated matching methods.

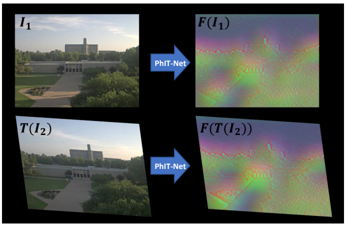


(a)

Patch size Method	32	64	128
QATM	0.568	0.761	0.853
DDIS	0.461	0.820	0.892
Original image	0.676	0.752	0.835
Lettry	0.733	0.805	0.853
Li-Snavely	0.712	0.806	0.888
Maddern	0.386	0.554	0.734
Ours	**0.781**	**0.888**	**0.930**

(b)

Figure 10: Patch matching results. (a) IoU-ROC curves and AUC scores, patch size 32. (b) AUC scores for all patch sizes. Experiment on 100 scenes, 10 patches in each scene.



(a)

Angle Method	2	10	18	26
Ground Truth	35.99	34.74	32.75	31.99
Original image	19.85	21.22	21.79	20.97
Lettry	21.13	21.20	17.98	15.26
Li-Snavely	26.51	28.22	27.86	27.22
Maddern	22.80	24.21	22.56	19.49
Ours	**28.71**	**30.10**	**30.04**	**29.53**

(b)

Figure 11: (a) Illustration of the registration test. Top left – Reference image (I_1); bottom left – Target image (I_2) transformed by a rigid transformation $T(\cdot)$; right – the respective photo-consistent transform. (b) PSNR results of the registration test.

Registration. The rigid registration test is performed based on two images (reference and target) of the same scene under different illumination. An affine transformation is applied to the target image (see an example in Fig. 11(a)). The goal of the registration algorithm is to estimate the reverse affine transformation matrix which aligns the transformed target and reference images. It is expected that an illumination invariant representation can improve the algorithm's accuracy. Registration is performed by ECC registration [9], based on cross correlation, as implemented in the OpenCV library. In Fig. 11 (right) the results for various angles are shown. The accuracy is measured by applying the estimated inverse transformation on the (transformed) target and computing the PSNR (with respect to the original target). Note that there are some minor errors also when the inverse transformation is known precisely (referred in the table as "Ground Truth"), due to numerical errors in applying the affine transformation. Our representation achieves the highest average PSNR.

6 Conclusion

A photo-consistent image transform is proposed, based on a data-driven invariant framework. The desired invariance property is learnt, while retaining geometrical coherence. We show that general and simple axioms yield state-of-the-art results, without resorting to over-simplified model constraints. Excellent matching and registration results are obtained by combining fast classical algorithms with our representation, also in extreme lighting variations. This idea can be generalized to design new representations, that are invariant to other types of nuisance image changes.

Acknowledgment

We thank Eyal Gofer for his help in improving the clarity and phrasing of the paper. We thank the research group members, Shachar, Ilya, Reut, Tom and Jonathan for their feedback on the paper. We acknowledge support by the Israel Science Foundation (Grant No. 534/19), the Ministry of Science and Technology (Grant 3-15621) and by the Ollendorff Minerva Center.

References

[1] Harry Barrow, J Tenenbaum, A Hanson, and E Riseman. Recovering intrinsic scene characteristics. *Comput. Vis. Syst*, 2(3-26):2, 1978.

[2] Harry G Barrow, Jay M Tenenbaum, Robert C Bolles, and Helen C Wolf. Parametric correspondence and chamfer matching: Two new techniques for image matching. Technical report, SRI INTERNATIONAL MENLO PARK CA ARTIFICIAL INTELLIGENCE CENTER, 1977.

[3] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):159, 2014.

[4] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[5] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a "siamese" time delay neural network. In *Advances in neural information processing systems*, pages 737–744, 1994.

[6] De Cheng, Yihong Gong, Sanping Zhou, Jinjun Wang, and Nanning Zheng. Person re-identification by multi-channel parts-based cnn with improved triplet loss function. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[7] Jiaxin Cheng, Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Qatm: quality-aware template matching for deep learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11553–11562, 2019.

[8] Michael Elad. Retinex by two bilateral filters. In *International Conference on Scale-Space Theories in Computer Vision*, pages 217–229. Springer, 2005.

[9] Georgios D Evangelidis and Emmanouil Z Psarakis. Parametric image alignment using enhanced correlation coefficient maximization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1858–1865, 2008.

[10] Graham D Finlayson, Mark S Drew, and Cheng Lu. Intrinsic images by entropy minimization. In *European conference on computer vision*, pages 582–595. Springer, 2004.

[11] Graham D Finlayson, Steven D Hordley, Cheng Lu, and Mark S Drew. On the removal of shadows from images. *IEEE transactions on pattern analysis and machine intelligence*, 28(1):59–68, 2005.

[12] Xueyang Fu, Delu Zeng, Yue Huang, Xiao-Ping Zhang, and Xinghao Ding. A weighted variational model for simultaneous reflectance and illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2782–2790, 2016.

[13] YZ Goh, Andrew BJ Teoh, and Michael KO Goh. Wavelet based illumination invariant preprocessing in face recognition. In *2008 Congress on Image and Signal Processing*, volume 3, pages 421–425. IEEE, 2008.

[14] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2335–2342. IEEE, 2009.

[15] David Hafner, Oliver Demetz, and Joachim Weickert. Why is the census transform good for robust optic flow computation? In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 210–221. Springer, 2013.

[16] Alexander Hermans, Lucas Beyer, and Bastian Leibe. In defense of the triplet loss for person re-identification. *arXiv preprint arXiv:1703.07737*, 2017.

[17] Elad Hoffer and Nir Ailon. Deep metric learning using triplet network. In *International Workshop on Similarity-Based Pattern Recognition*, pages 84–92. Springer, 2015.

[18] Ron Kimmel, Michael Elad, Doron Shaked, Renato Keshet, and Irwin Sobel. A variational framework for retinex. *International Journal of computer vision*, 52(1):7–23, 2003.

[19] BG Kumar, Gustavo Carneiro, Ian Reid, et al. Learning local image descriptors with deep siamese and triplet convolutional networks by minimising global loss functions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5385–5394, 2016.

[20] Edwin H Land and John J McCann. Lightness and retinex theory. *Josa*, 61(1):1–11, 1971.

[21] Louis Lettry, Kenneth Vanhoey, and Luc Van Gool. Unsupervised deep single-image intrinsic decomposition using illumination-varying image sequences. In *Computer Graphics Forum*, volume 37-7, pages 409–419. Wiley Online Library, 2018.

[22] Zhengqi Li and Noah Snavely. Learning intrinsic image decomposition from watching the world. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 9039–9048, 2018.

[23] Andrew Liu, Shiry Ginosar, Tinghui Zhou, Alexei A Efros, and Noah Snavely. Learning to factorize and relight a city. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 544–561. Springer, 2020.

[24] Wei-Chiu Ma, Hang Chu, Bolei Zhou, Raquel Urtasun, and Antonio Torralba. Single image intrinsic decomposition without a single intrinsic image. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 201–217, 2018.

[25] Will Maddern, Alex Stewart, Colin McManus, Ben Upcroft, Winston Churchill, and Paul Newman. Illumination invariant imaging: Applications in robust vision-based localisation, mapping and classification for autonomous vehicles. In *Proceedings of the Visual Place Recognition in Changing Environments Workshop, IEEE International Conference on Robotics and Automation (ICRA), Hong Kong, China*, volume 2, page 3, 2014.

[26] Laurence Meylan and Sabine Susstrunk. High dynamic range image rendering with a retinex-based adaptive filter. *IEEE Transactions on image processing*, 15(9):2820–2830, 2006.

[27] Thomas Müller, Clemens Rabe, Jens Rannacher, Uwe Franke, and Rudolf Mester. Illumination-robust dense optical flow using census signatures. In *Joint Pattern Recognition Symposium*, pages 236–245. Springer, 2011.

[28] Ana Belén Petro, Catalina Sbert, and Jean-Michel Morel. Multiscale retinex. *Image Processing On Line*, pages 71–88, 2014.

[29] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015.

[30] Daniel Scharstein, Heiko Hirschmüller, York Kitajima, Greg Krathwohl, Nera Nešić, Xi Wang, and Porter Westling. High-resolution stereo datasets with subpixel-accurate ground truth. In *German conference on pattern recognition*, pages 31–42. Springer, 2014.

[31] Moein Shakeri and Hong Zhang. Illumination invariant representation of natural images for visual place recognition. In *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 466–472. IEEE, 2016.

[32] John R Smith and Shih-Fu Chang. Transform features for texture classification and discrimination in large image databases. In *Proceedings of 1st International Conference on Image Processing*, volume 3, pages 407–411. IEEE, 1994.

[33] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[34] Itamar Talmi, Roey Mechrez, and Lihi Zelnik-Manor. Template matching with deformable diversity similarity. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 175–183, 2017.

[35] Jing-Wein Wang, Ngoc Tuyen Le, Jiann-Shu Lee, and Chou-Chen Wang. Illumination compensation for face recognition using adaptive singular value decomposition in the wavelet domain. *Information Sciences*, 435:69–93, 2018.

[36] Kilian Q Weinberger and Lawrence K Saul. Distance metric learning for large margin nearest neighbor classification. *Journal of Machine Learning Research*, 10(Feb):207–244, 2009.