

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

Supplementary Material

Damian Kaliroff
dkaliroff@technion.ac.il

Technion - Israel Institute of Technology
Haifa, Israel

Guy Gilboa
guy.gilboa@ee.technion.ac.il

In the supplementary material we provide details about the training of our model and explain in details about the training data. We also present additional evaluation examples of the proposed photo-consistent transform and we explain the ablation study performed during our research which resulted in the final configuration of our model.

1 Training Details

The hyperparameters values used in our final model are given below. In addition, we explain the implementation of the scale consistency loss.

1.1 Hyperparameters

- **Training steps:** Batch size=16, Epoch size=1000, Number of epochs=75.
- **Adam optimizer:** Momentum=0.9, Beta=0.999, Learning rate=1e-5.

1.2 Scale Consistency Loss

The scale consistency loss was defined in the paper, Eq. (11)

$$L_{SC}(f_a) = D_{scale}(F(G(f_a, \rho)), G(F(f_a), \rho)), \quad (1)$$

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 original patch size. It is calculated using an additional instance of PhIT-Net (apart from the A, P, N instances). The loss is computed only with respect to the anchor patches. The reason is that it is an "internal" loss, relating the patch to itself. It is not based on patch comparison, thus there is no need to duplicate it for the whole triplet allowing a faster training process. We remind that in the triplet network model the training is based on various instances of the same network with shared weights. Hence, a loss computed for one instance affects all instances.

2 Training and Test Data

2.1 Outdoors Dataset

We train and test our main model using outdoor images from the BigTime dataset (See Figure 1). Ideally, for training we would like the scenes to be completely static with changes only in illumination conditions. However, this is not always the case. Since the images are taken from time-lapse videos, there are small camera movements over time. Thus, alignment is not perfect. In addition, there are sky changes over time. Finally, the scenes are actually only semi-static, there are changes which happen over time such as cars or people that appear or disappear, windows that are open or shut, etc. Both the changing skies and the object changes are not part of the illumination variations we assume for the learning process. These issues were already raised by [9] who provide masks for regions which exhibit change not related to illumination. We take into consideration these masks in the patch selection at training. The camera movements are not corrected in the dataset. Thus we chose to manually select the most stable scenes (with minimal camera movement) for training.

The training is based on square patches of 64×64 pixels. This size is large enough to include most relevant semi-local illumination cues. We found that smaller patches do not contain enough details and the use of larger patches does not improve quality and considerably slows down the training. 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.



Figure 1: We use BigTime [9] as our outdoors dataset. For each scene there are several images under different lighting conditions. The dataset is composed of diverse scenes.

2.2 Indoors Dataset

In addition to our main outdoors dataset, we train and evaluate our model also on a set of indoor images. The dataset is comprised of 23 different scenes from Middlebury 2014 stereo dataset [10]). Each scene contains two images acquired under different lighting conditions. We train our model with patches extracted from 16 scenes (in the same manner done for BigTime) and test it with the remaining 7 scenes.

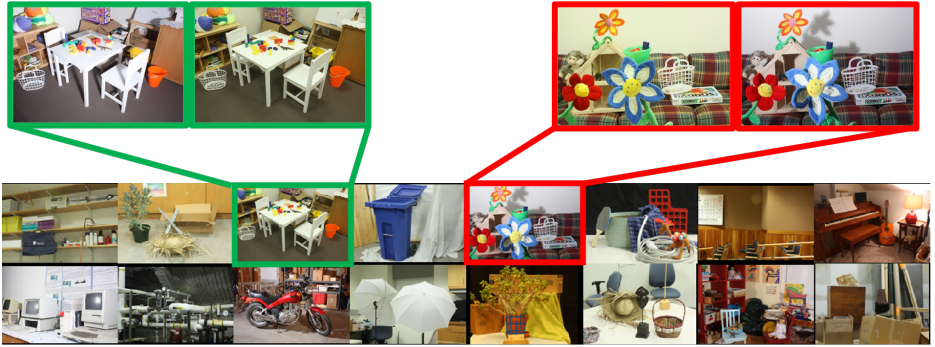


Figure 2: We use BigTime [8] as our outdoors dataset. For each scene there are several images under different lighting conditions. The dataset is composed of diverse scenes.

3 Additional Examples

3.1 Visual Photo-consistency

Indoors Dataset (Middlebury). As explained before, in addition to training a model with the BigTime dataset, we also trained and evaluated our model on an indoors dataset. In Fig. 3 we present additional examples of scenes from this dataset [9] with our transform, compared to other methods. We also show enlarged crops of interesting regions in the images. We can see in these examples that our representation, have the lowest differences compared to the others. We note that the Maddern representation of the bicycle scene (left) also exhibits a small difference between different illuminations. However, it removes significant structural information (and hence did not perform well in the quantitative experiments, as shown in the paper).

3.2 Patch Matching

Following Section 5.2 and Fig. 8 in the paper, we present in Fig. 4 additional matching results of difficult scenarios on both datasets. We show the matching results and the correlation-based heatmaps, according to which the algorithm selects its match.

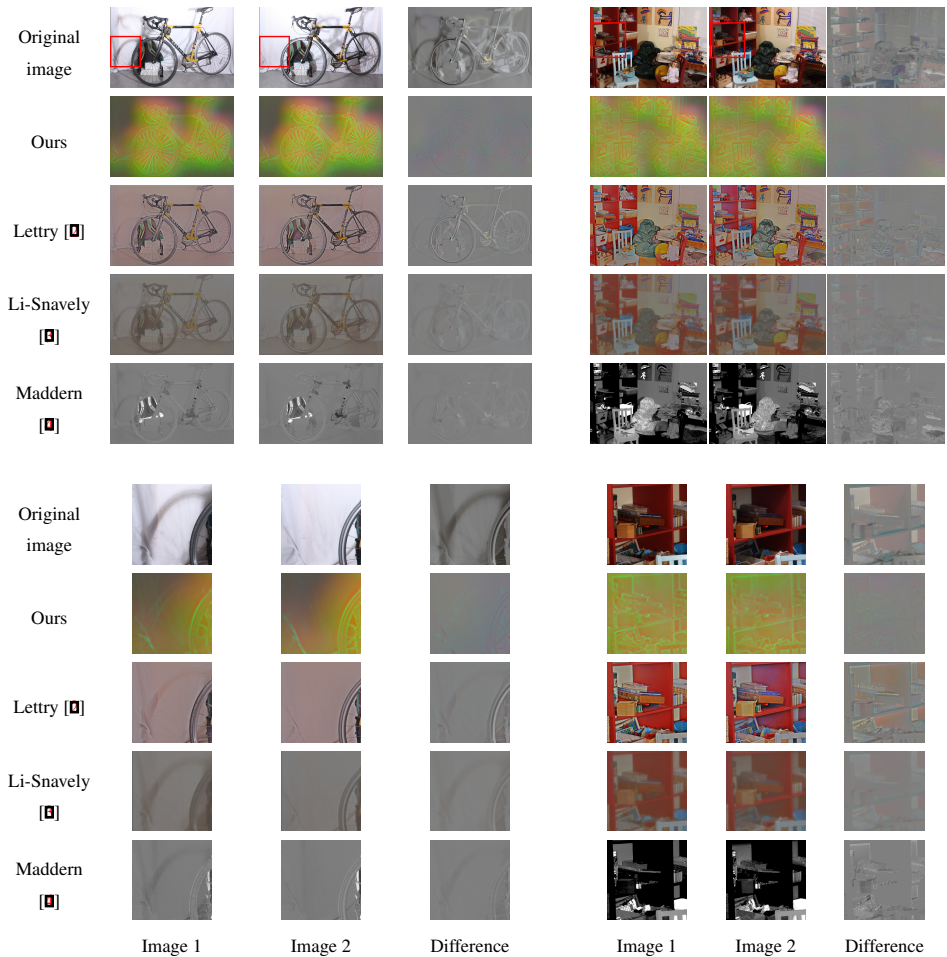


Figure 3: Visual comparison of invariant representation methods. For each scene, the representation of two images under different illumination conditions is shown. The difference (ideally zero) affirms that our representation is highly stable under illumination changes (zero is gray). In the top of the figure we see the comparison with the full images and in the bottom we show enlarged crops from the full images. The original crops location is marked with a red frame in the original images (first row).



Figure 4: Examples of patch matching results. The reference and target images belong to the same scene with different illumination. The goal is to correctly locate in the target image a patch which is selected in the reference image (green frame). The frames marking the results of the different algorithms are overlaid on the target image. Heatmaps of each algorithm (right) indicate high (red) to low (blue) matching scores. (Scenes above the line – BigTime, scenes below the line – Middlebury.)

4 Ablation Study

In this section we show different configurations of PhIT-Net. We examine two types of variations: Loss function variations and representations with different number of channels. The effects of these changes are quantified and visualized (Figure 5). The study is performed on the BigTime dataset by training a new network for each variation and evaluating its performance on the patch matching task, described in Section 5 in the paper.

4.1 Loss Function Variations

Scale Consistency Loss. In this study we set to zero the weight of the scale consistency loss,

$$L_{SC}(f_a) = D_{scale}(F(G(f_a, \rho)), G(F(f_a), \rho)), \quad (2)$$

Removing this loss reduces the sharpness of the representation, see Figure 5, column (c). This also affects the patch matching results. In the full model the performance is better for smaller patches (a matching task which is harder). However, the full model exhibits slightly worse accuracy for larger patches.

Multi-Channel Similarity Loss. In this study we set to zero the weight of the multi-channel similarity loss,

$$L_{MC}(I) = \sum_i \sum_{j \neq i} (1 - D_{corr}(I_i, I_j))^2. \quad (3)$$

Without this loss the channels of the representation tend to be similar to each other or the negative of each other (highly correlated or anti-correlated), see Figure 5, column (d). Adding this loss promotes variability amongst the different channels and reduce information loss.

Rotation invariance. Following the purpose of introducing the scale consistency loss, Eq. (2), it appears natural to introduce also a rotation consistency loss. This can be formalized as:

$$L_{RI}(f_a) = \|F(H(f_a, \rho)) - H(F(f_a), \rho)\|_2^2. \quad (4)$$

Where H rotates the image/representation by a random angle $\rho \in \{90, 180, 270\}$ degrees. This forces the representation to be invariant to (90 degree) rotations. Although this sounds highly reasonable (and might be necessary for some applications), we found out that the addition of this loss deteriorates performance. This might be explained by the fact that the color-coding of the dominant edge-direction, produced by the full model (see Section 5 in the paper), is direction dependant and thus it is lost here. (Figure 5, column (e)).

4.2 "K-Channel" Representation

Since our representation is unconstrained, in principle, it can be composed of an arbitrary number of channels. We trained and tested our full model with different number of output channels. This was achieved by changing the last convolutional layer of the network. We observe that 3 channels yield an optimal representation (in terms of matching).

<div>Patch Size</div> <div>Variation</div>	32	64	128
3 channels (Full model)	0.781	0.888	0.930
1 channel	0.700	0.847	0.907
2 channels	0.711	0.862	0.907
4 channels	0.779	0.873	0.924
5 channels	0.769	0.862	0.912
No scale consistency loss	0.740	0.891	0.941
No multi-channel similarity loss	0.720	0.838	0.893
With Rotation invariance loss	0.748	0.831	0.897

Table 1: Ablation study patch matching results: score by AUC of IoU-ROC curves.

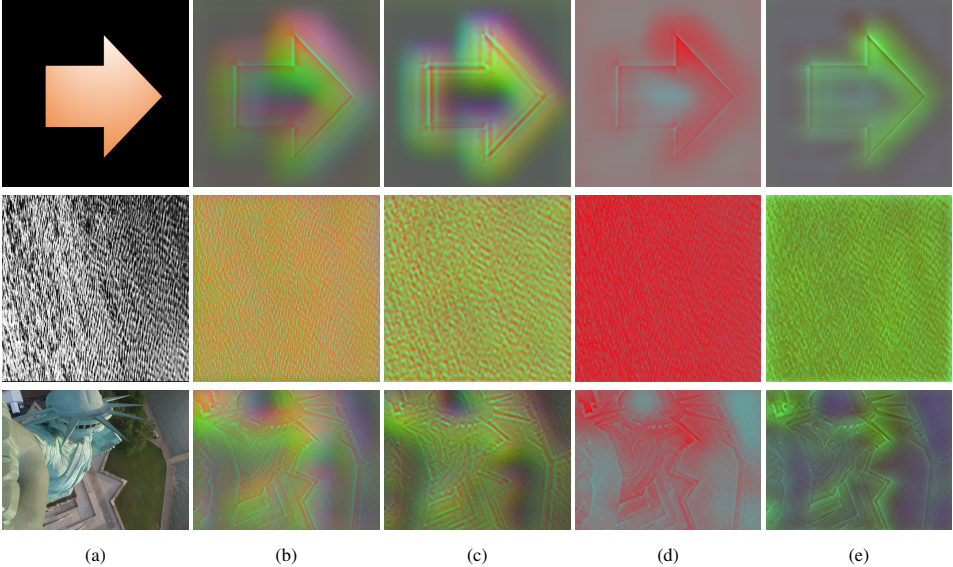


Figure 5: Variations of the representation. (a) Original image, (b) Full model representation, (c) No Scale consistency loss, (d) No Multi-channel similarity loss, (e) With Rotation invariance loss.

References

- [1] 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.
- [2] 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.
- [3] 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.
- [4] 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.
- [5] 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.