Learning Matchable Image Transformations for Long-term Metric Visual Localization

Lee Clement¹, Mona Gridseth², Justin Tomasi¹ and Jonathan Kelly¹

Abstract—Long-term metric self-localization is an essential capability of autonomous mobile robots, but remains challenging for vision-based systems due to appearance changes caused by lighting, weather, or seasonal variations. While experience-based mapping has proven to be an effective technique for bridging the ‘appearance gap,’ the number of experiences required for reliable metric localization over days or months can be very large, and methods for reducing the necessary number of experiences are needed for this approach to scale. Taking inspiration from color constancy theory, we learn a nonlinear RGB-to-grayscale mapping that explicitly maximizes the number of inlier feature matches for images captured under different lighting and weather conditions, and use it as a pre-processing step in a conventional single-experience localization pipeline to improve its robustness to appearance change. We train this mapping by approximating the target non-differentiable localization pipeline with a deep neural network, and find that incorporating a learned low-dimensional context feature can further improve cross-appearance feature matching. Using synthetic and real-world datasets, we demonstrate substantial improvements in localization performance across day-night cycles, enabling continuous metric localization over a 30-hour period using a single localization performance across day-night cycles, enabling continuous metric localization over a 30-hour period using a single

Index Terms—Deep Learning in Robotics and Automation, Visual Learning, Visual-Based Navigation, Localization

I. INTRODUCTION

SELF-LOCALIZATION is an essential capability of autonomous mobile robots, and localization algorithms based on inexpensive commercial vision sensors have become useful and widespread. Despite this success, long-term metric localization, where the goal is to continuously estimate the 6-dof pose of the vehicle with respect to a visual map, remains challenging in the presence of appearance change caused by illumination variations over the course of a day, changes in weather conditions, or seasonal shifts. This difficulty is largely due to simplifying assumptions such as brightness constancy and feature descriptor invariance that, when violated, cause visual localization systems to fail. Ideally, we would like our systems to function across this ‘appearance gap’, immune to variations in environmental conditions.

Manuscript received: September 10, 2019; Revised December 2, 2019; Accepted December 31, 2019.

This paper was recommended for publication by the Associate Editor and Reviewers’ comments.

¹Space and Terrestrial Autonomous Robotic Systems (STARS) Lab, University of Toronto Institute for Aerospace Studies (UTIAS), Canada.
²Autonomous Space Robotics Lab (ASRL), UTIAS, Canada.
<firstname>.<lastname>@robotics.utias.utoronto.ca.

Digital Object Identifier (DOI): see top of this page.

Long-term maps based on multiple visual ‘experiences’ of an environment have proven to be effective tools for metric localization through daily and seasonal appearance change [1]–[4]. In [3], consecutive visual experiences are recorded in a spatio-temporal pose graph, and localization against a privileged mapping experience proceeds by recalling a relevant experience and tracing through a chain of relative transformations in the graph. This process is often aided by a prior on the vehicle’s topological location in the graph, whether from dead reckoning, place recognition, or GNSS, which serves to limit the number of candidate vertices for metric localization. However, the number of intermediate ‘bridging’ experiences required for reliable long-term localization can be very large, and methods for compressing experience graphs are necessary for this approach to scale.

Recent work in [5], [6] has explored deep image-to-image translation [7], [8] as a means of directly bridging the ‘appearance gap’ and localizing with fewer experiences. However, these methods rely at least in part on well-aligned training images, which are difficult to obtain at scale in the real world. Moreover, the losses used to train these models are not explicitly connected to a target localization pipeline, and provide few assurances that the learned image transformations will ultimately improve localization performance.

We address these limitations by learning an image transformation optimized for a given combination of localization pipeline, sensor, and operating environment. Rather than translating between arbitrary appearance conditions, we learn to map images to a maximally matchable representation (i.e., one which maximizes the number of inlier feature matches) for a given feature detection/matching algorithm. Specifically, we learn a drop-in replacement for the standard RGB-to-grayscale colorspace mapping used to pre-process RGB images for use with conventional feature detection/matching algorithms, which typically operate on single-channel images (Figure 1). This formulation builds upon prior work on color constancy theory [9], does not require well-aligned images for training, and naturally admits a self-supervised training approach as training targets can be generated on the fly by the localization pipeline. Our main contributions are:

1) a technique for improving the robustness of a conventional visual localization pipeline to appearance change using a learned image pre-processing step;
2) a method for optimizing the performance of a non-differentiable localization pipeline by approximating the pipeline using a deep neural network;
3) experimental results on synthetic and real long-term vision datasets showing that our method enables con-
II. RELATED WORK

Appearance robustness in metric visual localization has previously been studied from the perspective of illumination invariance, with methods such as [11]–[14] making use of hand-engineered image transformations to improve feature matching over time for a given feature detector and descriptor. Similarly, affine models and other simple analytical transformations have been used to improve the robustness of direct visual localization to illumination change [15], [16]. Other approaches such as [17]–[20] have focused on learning feature descriptors that are robust to certain types of appearance change in autonomous route following applications. However, [17], [18] produce correspondences that are only weakly localized, and [19], [20] require sets of true and false point correspondences to train feature descriptors, which are challenging to obtain at scale over long periods.

Deep image-to-image translation [7], [8] has recently been applied to the problem of metric localization across appearance change. In [21] the authors train a convolutional encoder-decoder network to enhance the temporal consistency of image streams captured in environments with high dynamic range. Here the main source of appearance change is the camera itself as it automatically modulates its imaging parameters in response to the local brightness of a static environment. Other work has tackled the problem of localization across environmental appearance change, with [5] learning a many-to-one mapping onto a privileged appearance condition and [6] learning multiple pairwise mappings between appearance categories such as day and night. Image-to-image translation has also been applied to the related task of appearance-invariant place recognition [22], [23], which typically relies on patch matching or whole-image statistics to identify images corresponding to nearby physical locations rather than estimating the 6-dof pose of the vehicle. While [6], [21] include loss terms to maximize gradient information, these heuristics are not directly tied to the performance of the localization pipeline. Moreover, [5], [6], [21] require well-aligned training images exhibiting appearance variation, which are difficult to obtain at scale in the real world, and it is not clear how categorical appearance mappings such as [6], [22], [23] should be applied to continuous appearance change in long-term deployments.

Surrogate-based methods for approximating computationally expensive or non-differentiable objective functions are common in the numerical optimization literature [24]. Neural network surrogates in particular have found applications in a variety of domains including computer graphics [25] and computational oceanography [26], where high-fidelity physics simulations are available but expensive to compute. Our method of learning a differentiable loss function is similar in spirit to Generative Adversarial Networks (GANs) [27] in that a complex discriminator/loss function is trained using a comparatively simple analytical loss function. It also bears resemblances to perceptual losses [28], where the loss function is derived from the feature activations of a network trained on a proxy task such as image classification.

III. LEARNING MATCHABLE COLORSPACE TRANSFORMATIONS

Our goal in this work is to learn a nonlinear transformation \( f : \mathbb{R}^3 \rightarrow \mathbb{R} \) mapping the RGB colorspace onto a grayscale colorspace that explicitly maximizes a chosen performance metric of a vision-based localization pipeline. We investigate two approaches to formulating such a mapping: 1) a single function to be applied as a pre-processing step to all incoming images, similarly to [11], [13], [14]; and 2) a parametrized function tailored to the specific image pair to be used for localization, where the parameters of this function are derived from the images themselves. Additionally, the functional form of either mapping may be specified analytically (e.g., from physics) or learned from data using a function approximator such as a neural network.

In order to find an optimal colorspace transformation for a given application, we require an appropriate objective function to optimize, which should ideally be tied to the performance of the target localization pipeline. An intuitive choice of objective

1.github.com/utiasSTARS/matchable-image-transforms
In this work we based optimization schemes commonly used in deep learning, and RANSAC [29], which are incompatible with the gradient-mance metric to optimize, the most commonly used local-adopt the latter approach in this work, since high-quality 6-dof ground truth is difficult to obtain over long time scales.

Although it is straightforward to choose a target performance metric to optimize, the most commonly used localization front-ends in robotics rely on non-differentiable components such as stereo matching, nearest-neighbors search, and RANSAC [29], which are incompatible with the gradient-based optimization schemes commonly used in deep learning. In this work we learn an objective function by training a deep convolutional neural network (CNN) to act as a differentiable proxy to the localization front-end. Specifically, we train a siamese CNN to predict the number of inlier feature matches for a given image pair, where the training targets are generated using a conventional non-differentiable feature detector/matcher algorithm based on libviso2 features [30]. This proxy network can then be used to define a fully differentiable objective function, allowing us to train a nonlinear colorspace mapping using gradient-based methods. Finally, the trained image transformation can be used as a pre-processing step in a conventional visual localization pipeline, enabling it to operate more reliably under appearance change. Figure 1 summarizes our full data pipeline pictorially.

A. Differentiable Matcher Proxy

We consider the task of training a CNN $M_\theta$, with parameters $\theta$, to predict the number of inlier feature matches returned by a non-differentiable feature detector/matcher $M$ for a given image pair $(I_1, I_2)$. This training objective is a convenient choice for our intended application as it is closely tied to the ability of our visual localization pipeline to operate across appearance change, however it is not by any means the only choice. For example, we could also train a CNN to predict a measure of localization accuracy such as geodesic distance from ground truth on the SE(3) manifold, similarly to the estimator correction framework proposed in [31]. Importantly, this formulation admits a self-supervised training approach as this formulation admits a self-supervised training approach as

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left( M_\theta(I_1^i, I_2^i) - M(I_1^i, I_2^i) \right)^2.$$  \hspace{1cm} (1)

B. Physically Motivated Transformations

Prior work in [9] has shown that under the assumptions of a single black-body illuminant and an infinitely narrow sensor response function, an appropriately weighted linear combination of the log-responses of a three-channel (e.g., RGB) camera represents a projection onto an invariant one-dimensional chromaticity space that is independent of both the intensity and color temperature of the illuminant, and depends only on the imaging sensor and the materials in the scene:

$$F_j^i = \log I_j^i(\lambda_2) - \alpha \log I_j^i(\lambda_1) - \beta \log I_j^i(\lambda_3),$$  \hspace{1cm} (2)

where $I_j^i(\lambda_k)$ is the image of sensor responses at wavelength $\lambda_k$, the weights $\alpha, \beta$ are subject to the constraints

$$\frac{1}{\lambda_2} = \frac{\alpha}{\lambda_1} + \frac{\beta}{\lambda_3} \quad \text{and} \quad \beta = (1 - \alpha),$$  \hspace{1cm} (3)

and the indices $k$ are chosen such that $\lambda_1 < \lambda_2 < \lambda_3$ (i.e., red, green and blue channels, respectively). The image formed from this pixel-wise linear combination of log-responses can then be rescaled to produce a valid

![Fig. 2: Network architecture for $M_\theta$. Each block is denoted by stride, kernel size, input channels, and output channels. The left and right branches share weights. The final output is produced by a fully-connected layer (implemented as a convolution operator), which projects the feature map to a scalar value.](image-url)
This rescaling scheme allows the model to saturate parts of
the output images while still using the full range of valid pixel
values. Moreover, it avoids introducing significant sparsity in
the gradients through the \( \min(\cdot) \) and \( \max(\cdot) \) operators, which
improves the flow of gradient information during training.

### C. Learned Nonlinear Transformations

While the assumption of a single black-body illuminant in
[9] is reasonable for daytime navigation where the dominant
light source is the sun, it does not hold in many common
navigation scenarios such as nighttime driving with headlights.
Moreover, the assumption of an infinitely narrow sensor re-
response is unrealistic for real cameras. As an alternative to
the physically motivated colorspace transformation outlined
in Section III-B, we investigate the possibility of learning a
bespoke nonlinear mapping that maximizes matchability for
a particular combination of imaging sensor, estimator and
environment. We parametrize this mapping using a small neu-
ral network \( T_\psi \), with parameters \( \psi \), operating independently
on each pixel of each input RGB image. We structure \( T_\psi 
\) as a multilayer perceptron (MLP) implemented using \( 1 \times 1 
\) convolutions and PReLU nonlinearities.

We consider two versions of this MLP-based transformation,
both with and without incorporating an additional pairwise
context feature obtained from encoder network \( E_\phi \) using
Equation (5). In the case where \( E_\phi \) is used, the input to
\( T_\psi \) becomes the concatenation of the input RGB image and
the parameters \( \eta^i \) along the channel dimension, and the first
convolutional layer of \( T_\psi \) is modified accordingly. We train \( T_\psi 
\) and \( E_\phi \) (if used) jointly by minimizing a similar loss function
to Equation (6), where in place of Equation (4), we have

\[
F_j^i = \begin{cases} 
T_\psi(\eta^i) & \text{if } E_\phi \text{ is used,} \\
T_\psi(\eta) & \text{if } E_\phi \text{ is not used.} 
\end{cases}
\]

Similarly to the physically motivated transformations de-
scribed in Section III-B, we rescale \( F_j^i \) to fill the range of
valid grayscale values by applying Equation (7).

### IV. Experiments

We conducted experiments on synthetic and real-world
long-term vision datasets to validate and compare each ap-
proach. Specifically, we evaluated the ability of our matcher
proxy network to capture the performance of libviso2
feature matching across viewpoint and appearance changes,
as well as the effect of each image transformation on feature
matching and localization performance. When evaluating fea-
ture matching, we assumed that we had a prior on the vehicle’s
topological location in the map, such that we could reliably
identify the nearest vertex in the pose graph. This is typical
for autonomous visual route-following systems such as [4],
where the topological prior is derived by dead reckoning from
a previous successful localization, or using place recognition
or GNSS in the event that the system becomes lost.

We refer to the generalized color-constancy model of [9]
(Section III-B) as “SumLog” and “SumLog-E”, where the
latter uses Equation (5) to derive the parameters \( \eta^i \) per image
pair, and the former uses a constant \( \eta \) that maximizes inlier
feature matches over the training set (similarly to [11], [14]).
Fig. 3: Estimated vs. actual match counts for $M_\theta$ after ten epochs of pre-training on each dataset, colour-coded by relative density. Match counts are aggregated over all test sequences and include self-matches. The dashed red line corresponds to perfect agreement of $M_\theta$ with libviso2. In each case, the match count predictions produced by $M_\theta$ are very strongly correlated with the true match counts.

Analogously, we refer to the learned multilayer perceptron models (Section III-C) as “MLP” and “MLP-E”, where the latter incorporates $E_\phi$ and the former does not. We refer to the standard RGB-to-grayscale transformation as “Gray”.2

Training proceeds in two stages. First, we pre-train $M_\theta$ using standard grayscale images. Training labels are generated using the open-source libviso2 library [30] to detect and match features, and the eight-point RANSAC algorithm to reject outlier matches. Second, we train $E_\phi$ and/or $T_\psi$ using the matchability loss defined in Equation (6). To ensure that $M_\theta$ accurately predicts feature match counts for the output images, which differ significantly from standard grayscale images, we continue to train $M_\theta$ in an alternating fashion using the output images at each iteration. All models are implemented in PyTorch [10] and trained for 10 epochs with a batch size of 8, using the Adam optimizer [37] with default parameters and a learning rate of $10^{-4}$. We rescale all images to a height of 192 pixels for both training and testing.

A. Datasets

We evaluated our approach using both synthetic and real-world datasets exhibiting severe illumination change.

1) Virtual KITTI: The Virtual KITTI (VKITTI) dataset [38] is a synthetic reconstruction of a portion of the KITTI vision benchmark [39], consisting of five sets of non-overlapping trajectories with RGB-D imagery rendered under a variety of simulated illumination conditions. This dataset is a convenient validation tool as it provides perfect data

2We refer specifically to the ITU-R 601-2 luma transform implemented by the Pillow library: $L = 0.299R + 0.587G + 0.114B$. 
association and a range of daytime illumination conditions. For each trajectory, we train models using image pairs from the others. Since each trajectory is non-overlapping, this spatial split allows us to evaluate how well our method generalizes to unseen environments. Further, since VKITTI provides corresponding images from different viewpoints, we augmented the training data to ensure generalizability to viewpoint changes and dynamic objects by associating each training image in one condition with a window of nearby images in the other.

b) UTIAS In The Dark: The UTIAS In The Dark (InTheDark) dataset [3] provides stereo imagery of a 250 m outdoor loop traversed repeatedly over a 30-hour period using on-board headlights to illuminate the scene at night. We use the multi-instance localization system in [3] to obtain corresponding image pairs with overlapping fields of view but non-identical poses. We train our models using left-camera images from 66 traversals and test on 7 held-out traversals spanning a full day-night cycle (listed in Table I). This temporal split allows us to evaluate how well our method generalizes to multiple unseen illumination conditions.

c) Oxford RobotCar: We further evaluate our method using three sequences from the Oxford RobotCar (RobotCar) dataset [40], captured along the same 10 km route in overcast, nighttime, and sunny conditions.3 We find corresponding images using the GNSS/INS poses, and split the trajectory into two non-overlapping segments: the first 70% of the trajectory is used for training and the remaining 30% is used for testing.

\[ \text{Inlier Feature Matches} = \theta \]

\[ \text{Actual inlier feature matches using libviso2 and each RGB-to-grayscale transformation. The highest mean number of matches for each sequence is highlighted in bold.} \]

| Test Sequence     | Gray | SumLog | SumLog-E | MLP  | MLP-E |
|-------------------|------|--------|----------|------|-------|
| VKITTI/0001       | 262 (82) | 726 (136) | 689 (157) | 661 (108) | 623 (116) |
| Sunset-Morning    | 444 (58) | 790 (129) | 767 (125) | 747 (106) | 770 (107) |
| Overcast-Clone    | 240 (22) | 812 (67) | 803 (70)  | 774 (65)  | 702 (62)  |
| VKITTI/0002       | 290 (41) | 739 (67) | 757 (76)  | 755 (65)  | 764 (84)  |
| Sunset-Morning    | 125 (33) | 669 (44) | 735 (43)  | 711 (37)  | 642 (41)  |
| Overcast-Clone    | 142 (33) | 647 (39) | 666 (43)  | 566 (47)  | 456 (38)  |
| VKITTI/0006       | 234 (53) | 817 (36) | 816 (37)  | 745 (37)  | 731 (40)  |
| Sunset-Morning    | 311 (46) | 548 (52) | 555 (50)  | 450 (42)  | 486 (39)  |
| Overcast-Clone    | 210 (39) | 762 (69) | 758 (71)  | 756 (62)  | 675 (69)  |
| VKITTI/0020       | 287 (78) | 716 (71) | 716 (72)  | 718 (62)  | 708 (64)  |
| Sunset-Morning    | 11(3)    | 12 (2)   | 12 (3)    | 11 (2)    | 11 (2)    |
| Overcast-Sunny    | 26 (12)  | 25 (12)  | 25 (12)   | 24 (9)    | 71 (41)   |

B. Feature Matcher Approximation

We train $\mathcal{M}_\theta$ in a self-supervised manner to predict the number of inlier feature matches for overlapping image pairs captured under different illumination conditions from nearby but different poses. Training labels are generated on the fly for each image pair using the open-source libviso2 library [30] in monocular flow matching mode with default parameters, using the eight-point RANSAC algorithm to reject outlier matches. In practice we train $\mathcal{M}_\theta$ to minimize Equation (1) over all combinations of input images.

Figure 3 plots actual and estimated match counts for each dataset after ten epochs of pre-training on Gray images, aggregated over all test sequences. These include self-matches (same viewpoint, same appearance) and non-self-matches (different viewpoint, different appearance), which appear as clusters. In each case the test-time match counts predicted by $\mathcal{M}_\theta$ are strongly correlated with the true performance of libviso2. This indicates that our approach generalizes well and that $\mathcal{M}_\theta$ is capturing salient properties of feature matching rather than memorizing training examples.

C. Feature Matching Across Appearance Change

Figure 4 shows the outputs of each image transformation for sample RGB image pairs in the VKITTI/0020 Morning and Sunset sequences (Figure 4a) and the challenging sequence InTheDark/0041 (Figure 4b). We see that each model produced image pairs that are visually more consistent than standard Gray images, and that local illumination variations

3“Overcast”, “Night”, and “Sunny” refer to 2014-12-09-13-21-02, 2014-12-10-18-10-50, and 2014-12-16-09-14-09, respectively.
TABLE II: Maximum distances travelled on dead reckoning for each test sequence of the UTIAS In The Dark dataset, based on various thresholds of inlier feature matches against the “Map” sequence. The best results are highlighted in bold.

| Sequence   | 10 Inliers | 20 Inliers | 30 Inliers |
|------------|------------|------------|------------|
| InTheDark  | G1 S2 S-E2 M1 M-E3 | G1 S2 S-E2 M1 M-E3 | G1 S2 S-E2 M1 M-E3 |
| Map (08:54) | - - - - - | - - - - - | - - - - - |
| 0006 (09:46) | 0.0 0.0 0.0 0.0 0.0 | 0.0 0.0 0.0 0.0 0.0 | 0.0 0.1 0.0 0.0 0.0 |
| 0027 (18:36) | 7.5 0.7 0.0 0.3 0.3 | 25.9 3.6 0.7 2.5 2.2 | 104.5 15.1 10.0 18.6 15.7 |
| 0041 (21:48) | 14.9 0.5 0.2 0.9 1.4 | 46.3 8.4 3.2 7.2 7.4 | 104.5 15.1 10.0 18.6 15.7 |
| 0058 (05:48) | 0.0 0.0 0.0 0.0 0.0 | 0.0 0.2 0.0 0.2 0.0 | 0.0 0.6 0.3 0.6 0.0 |
| 0071 (09:18) | 0.0 0.0 0.0 0.0 0.0 | 0.0 0.0 0.0 0.0 0.0 | 0.0 0.0 0.0 0.0 0.0 |
| 0083 (14:01) | 0.0 0.2 0.0 0.0 0.0 | 0.4 0.5 0.3 0.2 0.0 | 2.2 4.0 0.3 0.4 0.5 |
| 0089 (16:43) | 0.3 0.3 0.2 0.0 0.0 | 4.8 1.0 3.7 1.2 0.8 | 6.1 6.1 4.7 2.9 2.9 |
| RobotCar   | Night 27.0 7.6 3.7 27.3 27.3 | 307.1 307.1 245.9 400.8 398.9 | 740.5 541.5 740.5 697.3 740.5 |
| Sunny      | 1.3 1.3 1.5 1.8 0.9 | 36.1 65.1 48.0 39.4 63.0 | 129.1 139.8 109.2 124.3 165.0 |

1 G: Gray  2 S: SumLog  3 M: MLP

such as shadows, uneven lighting, and specular reflections were minimized by optimizing Equation (6).

Figure 5 visually compares the distributions of actual libviso2 feature matches for each transformation, while Table I summarizes the results numerically. Each model significantly increased the mean number of inlier matches across most test sequences, with the greatest improvements generally obtained from the SumLog and SumLog-E transformations. Sequences InTheDark/0006 and InTheDark/0071 are exceptions in that the standard Gray transformation performed best. These sequences were recorded under similar conditions to the “Map” sequence, so feature matching can be expected to perform optimally on Gray images. We saw little improvement in match counts on the RobotCar/Overcast-Night experiment, which we attribute to motion blur in the nighttime images making feature matching exceptionally difficult. In contrast, the MLP-E method more than doubled the mean number of feature matches in the Sunny experiment.

We note that the pairwise encoder did not confer any significant benefit on the VKITTI sequences. These results are consistent with [11]–[14], where one or two sets of parameter values were sufficient to achieve good performance across varying daytime conditions. In contrast, the encoder network provided a noticeable performance boost on most InTheDark and RobotCar sequences. We attribute this difference to more variation in illumination and terrain, as a single transformation is less likely to perform well under more varied conditions. We also note that the MLP-E transformation frequently performed similarly to the SumLog-E transformation, suggesting that, in spite of key assumptions being broken, a linear combination of log-responses as proposed by [9] may in fact be an optimal solution for this problem, and that a careful choice of weights is the key to obtaining good cross-appearance feature matching over day-night cycles.

D. Impact on Localization Performance

We evaluate localization performance in an autonomous route-following context by examining the maximum distances in each sequence that would have been navigated using dead reckoning (e.g., visual odometry) as a result of failing to localize against the map. These results are summarized in Table II for thresholds of 10, 20, and 30 inlier feature matches against the “Map” sequence. A typical criterion for requiring manual intervention is dead reckoning in excess of 10 meters, depending on the accuracy of the underlying dead reckoning system. Based on a relatively conservative threshold of 20 inlier feature matches against the “Map” sequence, we see that InTheDark/0027 (evening) and InTheDark/0041 (night) presented significant difficulty for localization, which was substantially alleviated using any of the four image transformations. In particular, the SumLog-E transformation yielded maximum dead reckoning distances below four meters across all illumination conditions. We see similar improvements using the SumLog-E method with a threshold of 10 inliers on the RobotCar dataset. Together, these results imply that near-continuous 6-dof visual localization over a full day-night cycle is achievable using only a single mapping experience and a simple image pre-processing step, representing a dramatic reduction in data requirements to scale experience-based localization to long deployments.

With more conservative thresholds (e.g., 30 inliers), localization failures became more common, as expected. However, the proposed image transformations continued to provide substantially more robust localization performance compared to standard grayscale images. For example, we achieved a maximum distance on dead reckoning of 10.0 m using the SumLog-E method on sequence InTheDark/0041 with a threshold of 30 inliers. In contrast, using the standard Gray method would have required the system to rely continuously on dead reckoning for approximately 40% of the route.

V. Conclusions and Future Work

This paper presented a method for learning pixel-wise RGB-to-grayscale colorspace mappings that explicitly maximize the number of inlier feature matches for a given input image pair, feature detector/matcher and operating environment. Our key insight is that by training a deep neural network to predict the performance of a conventional non-differentiable feature detector/matcher, we can define a fully differentiable loss function that can be used to learn image transformations optimized
for localization performance. We evaluated our approach using both physically motivated and data-driven transformations and demonstrated substantially improved feature matching and localization performance on synthetic and real long-term vision datasets exhibiting severe illumination change, allowing experience-based localization to scale to long deployments with dramatically fewer bridging experiences. We consistently achieved the best performance using a physically motivated weighted sum of log-responses with weights derived from a pairwise context encoder network. In future work we plan to explore alternative loss functions such as pose estimation error, the use of feature locations or photometric consistency as a more granular supervision signal, and the impact of context feature dimension on MLP-based transformations. We also intend to investigate the impact of our method on feature matching across seasonal appearance change.

REFERENCES

[1] W. Churchill and P. Newman, “Experience-based navigation for long-term localisation,” Int. J. Robot. Res., vol. 32, pp. 1645–1661, Dec. 2013.

[2] C. Linegar, W. Churchill, and P. Newman, “Work smart, not hard: Recalling relevant experiences for vast-scale but time-constrained localisation,” in Proc. IEEE Int. Conf. Robot. Automat. (ICRA), pp. 90–97, May 2015.

[3] M. Paton, K. MacTavish, M. Warren, and T. D. Barfoot, “Bridging the appearance gap: Multi-experience localisation for long-term visual teach and repeat,” in Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS), pp. 1918–1925, Oct. 2016.

[4] M. Paton, K. MacTavish, L.-P. Berzci, S. K. van Es, and T. D. Barfoot, “I can see for miles and miles: An extended field test of visual teach and repeat 2.0,” in Proc. Field and Service Robot. (FSR), pp. 415–431, 2018.

[5] L. Clement and J. Kelly, “How to train a CAT: Learning canonical appearance transformations for direct visual localisation under illumination change,” IEEE Robot. Automat. Lett., vol. 3, no. 3, pp. 2447–2454, 2018.

[6] H. Porav, W. Maddern, and P. Newman, “Adversarial training for adverse conditions: Robust metric localisation using appearance transfer,” in Proc. IEEE Int. Conf. Robot. Automat. (ICRA), pp. 1011–1018, May 2018.

[7] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-image translation with conditional adversarial networks,” in Proc. IEEE Conf. Comput. Vision and Pattern Recognit. (CVPR), pp. 5967–5976, July 2017.

[8] J. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using Cycle-Consistent adversarial networks,” in Proc. IEEE Int. Conf. Comput. Vision (ICCV), pp. 2242–2251, Oct. 2017.

[9] S. Ramasingam and S. Collins, “Study of the photodetector characteristics of a camera for color constancy in natural scenes,” J. Optical Society of America A, vol. 27, no. 2, pp. 286–294, 2010.

[10] A. Paszek, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, “Automatic differentiation in PyTorch,” in Proc. NeurIPS Autodiff Workshop, 2017.

[11] L. Clement, J. Kelly, and T. D. Barfoot, “Robust monocular visual teach and repeat aided by local ground planarity and color-constant imagery,” J. Field Robot., vol. 34, no. 1, pp. 74–97, 2017.

[12] P. Corke, R. Paul, W. Churchill, and P. Newman, “Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation,” in Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS), pp. 2085–2092, 2013.

[13] C. McManus, W. Churchill, W. Maddern, A. D. Stewart, and P. Newman, “Shady dealings: Robust, long-term visual localisation using illumination invariance,” in Proc. IEEE Int. Conf. Robot. Automat. (ICRA), pp. 901–906, May 2014.

[14] M. Paton, F. Pommerleau, K. MacTavish, C. J. Ostafew, and T. D. Barfoot, “Expanding the limits of vision-based localization for long-term route-following autonomy,” J. Field Robot., vol. 34, no. 1, pp. 98–122, 2017.

[15] J. Engel, J. Stueckler, and D. Cremers, “Large-scale direct SLAM with stereo cameras,” in Proc. IEEE/RSJ Int. Conf. Intell. Robot. Syst. (IROS), pp. 1935–1942, 2015.

[16] S. Park, T. Schöps, and M. Pollefeys, “Illumination change robustness in direct visual SLAM,” in Proc. IEEE Int. Conf. Robot. Automat. (ICRA), pp. 4523–4530, May 2017.