Zero-Shot Day-Night Domain Adaptation with a Physics Prior

Attila Lengyel[1]   Sourav Garg[2]   Michael Milford[2]   Jan C. van Gemert[1]

Delft University of Technology[1]   QUT Centre for Robotics[2]

{a.lengyel, j.c.vangemert}@tudelft.nl   {s.garg, michael.milford}@qut.edu.au

Abstract

We explore the zero-shot setting for day-night domain adaptation. The traditional domain adaptation setting is to train on one domain and adapt to the target domain by exploiting unlabeled data samples from the test set. As gathering relevant test data is expensive and sometimes even impossible, we remove any reliance on test data imagery and instead exploit a visual inductive prior derived from physics-based reflection models for domain adaptation. We cast a number of color invariant edge detectors as trainable layers in a convolutional neural network and evaluate their robustness to illumination changes. We show that the color invariant layer reduces the day-night distribution shift in feature map activations throughout the network. We demonstrate improved performance for zero-shot day to night domain adaptation on both synthetic as well as natural datasets in various tasks, including classification, segmentation and place recognition.

1. Introduction

Deep image recognition methods are sensitive to illumination shifts caused by accidental recording conditions such as camera viewpoint, light color, and illumination changes caused by time of day or weather [1, 16, 78], as for example a model trained with daylight will not generalize to night time. Robustness to such recording conditions is essential for autonomous driving and other safety-critical computer vision applications. An illumination shift between train and test data is typically addressed by unsupervised domain adaptation [55, 57, 76] where the labeled training set is from one domain and the test set is from a different domain. The main assumption is that the test data is readily available and the challenge is how to make use of the unlabeled test data in an unsupervised setting to address the domain shift. However, adding test data is often non-trivial as it may be expensive and time consuming to obtain and due to the long tail of the real world is impossible to collect for all possible scenarios in advance.
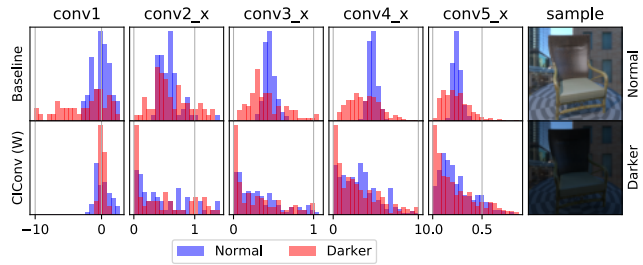
Instead of adding more data, prior knowledge can be

Figure 1: Feature map activations in various layers of a baseline ResNet-18 and a color invariant W-ResNet-18, averaged over all samples in a 'Normal' and 'Darker' test set (samples on right). The intensity change between the test sets causes an internal distribution shift throughout all layers of the baseline model. W normalizes the input resulting in more domain invariant features.

built-in as a visual inductive bias. The champion of such a bias is the convolution operator added to a deep network which yields a Convolutional Neural Network (CNN). The CNN is translation invariant, and thus saves a massive amount of data as the deep network no longer needs training samples at all possible locations. Here, we replace data by an inductive photometric bias. We introduce a novel zero-shot domain adaptation method for addressing day-night domain shifts exploiting learnable photometric invariant features as a physics-based visual inductive prior. In contrast to unsupervised domain adaptation, our zero-shot method reduces the data dependency by removing any reliance on the availability of test data.

Illumination changes to the source domain induce a distribution shift of feature map activations throughout all layers of a CNN. This is shown as the baseline in the top row of Fig. 1, where the activations of a CNN trained on daytime data are shown for a 'Normal' (source) and 'Darker' (target) test set. Such a distribution shift, in turn, has a severe detrimental effect on the accuracy of the CNN [39]. Because the distribution shift is between the training data and unavailable test data, this shift cannot be addressed in a data-driven manner using, for example, variants of Batch Normalization [29, 39]. Instead, we normalize feature map

activations in a data-free setting by exploiting photometric invariant features which are explicitly designed to tackle distribution shifts caused by illumination changes.

Photometric invariant features, or color invariants, represent object properties irrespective of the accidental recording conditions [24, 25], including 1) scene geometry, which affects the formation of shadows and shading, the 2) color and 3) intensity of the light source, which changes the overall tint and brightness of the scene, and 4) Fresnel reflections occurring on shiny materials where the incoming light is directly reflected from the surface without interacting with the material color. Thanks to their robustness to these lighting changes, color invariants have been widely used in classical computer vision applications [6, 46], yet their use in a deep learning setting has remained largely unexplored. We implement the color invariant edge detectors from [24] as a trainable Color Invariant Convolution (CIConv) layer which can be used as the input layer to any CNN to transform the input to a domain invariant representation. Fig. 1, bottom row, shows that CIConv reduces the distribution shift between the source and target test set in all network layers, improving target domain performance.

We have the following contributions: (i) we introduce CIConv, a learnable color invariant CNN layer that reduces the activation distribution shift in a CNN under an illumination-based domain shift; (ii) we evaluate several color invariants in the day-night domain adaptation setting on our two carefully curated classification datasets; and (iii) we demonstrate performance improvements on tasks related to autonomous driving, including classification, segmentation and place recognition. All datasets and code will be made available on our project page.[1]

2. Related work

Domain Adaptation   The aim of domain adaptation [76] is to train a model on a source domain dataset such that it performs well on a different but similar target domain dataset. This alleviates the burden of annotating datasets for applications in new domains where insufficient training data is available. Popular approaches rely on generative adversarial networks (GANs) to generate synthetic target domain samples [27] or aim to minimize the feature divergence between the two domains through an adversarial term [28, 65] or a discrepancy metric [66, 45] in the loss function. The day-night domain adaptation setting is particularly important due to the promise of self-driving cars and thus includes much work [13, 16, 18, 55, 56, 57, 58, 69, 73, 78] for semantic segmentation, and [7, 32, 51] for place recognition. However, all aforementioned methods (except [13]) require either training data from the target domain or additional modalities, whereas our approach uses only source

domain image data. Our approach requires no extra information sources and thus preempts expensive data gathering costs.

Zero-shot Domain Adaptation   Research on zero-shot learning [2, 37, 48, 49, 79, 83] has been readily extended from unseen classes to unseen domains, where domain adaptation is performed without having access to the target domain. However, current zero-shot domain adaptation methods require additional information in the form of: (i) extra task-irrelevant source and target domain data pairs to adapt to the task-relevant target domain [50, 75]; (ii) a parametrization of the domain shift by an attribute, where the attribute probability distribution for the unseen target domain is required to be known [30]; (iii) additional data from domains besides the source and target domain to learn a domain-invariant subspace projection [80], or; (iv) extra data in a partially labelled target domain [77]. These four types of information are generally not known for day-night domain shifts and are therefore not directly applicable. AdaBN [39] argues that domain-specific knowledge is stored in the batch normalization (BN) [29] layers of a model and performs domain adaptation by resampling BN statistics from the target domain. This again requires access to the target domain dataset. AdaBN [39] can be considered zero-shot if only the statistics of the current batch are used. However, this makes the method reliant on large batch sizes where classes are evenly represented. In contrast, our method does not require any information from the target domain other than the task agnostic physics-based illumination prior given by color invariants which are readily available from literature.

Physics-Guided Neural Networks   Adding prior knowledge from physical models in a neural network has the potential to improve performance without additional training data. The canonical example is adding translation equivariance through a convolutional prior [33, 68] where recent work shows benefits from adding prior knowledge, for example in line detection [43], spectral leakage [61] and antialiasing in CNNs [82]. In the case of physical image formation models, recent examples include intrinsic image decomposition [10], underwater image enhancement [84], or rain image restoration [38]. Here, we add an physical image formation prior to compensate for the lack of data in zero-shot domain adaptation. We investigate a relatively unexplored direction combining deep learning with physical color and reflection invariants.

Color invariants   The use of physics-based reflection models to improve invariance to illumination changes is a well-researched topic in classical computer vision [8, 11, 23, 25, 70, 71, 72]. Early work includes invariants derived

---

from the Kubelka-Munk (KM) reflection model [36, 24]. Based on the image formation model introduced in [20] various methods have been proposed for shadow removal or intrinsic image decomposition [19, 21] with applications in place recognition [15, 46], road detection [5, 6, 34, 35] and street image segmentation [67]. Recent works have shown improved segmentation performance by applying a color invariant transformation as a preprocessing step [3, 4, 47] or using the ground truth albedo as input on a synthetic dataset [9]. [1] demonstrates the sensitivity of CNNs to changes in white balance (WB) settings and shows how robustness can be improved using an auto-WB preprocessing step. Our work further explores the use of classical color invariants as a trainable deep network layer.

3. Method

Our color invariant layers make use of the invariant edge detectors from [24]. The edge detectors are derived from the image formation model based on the Kubelka-Munk theory [36] for material reflections, which describes the spectrum of light E reflected from an object in the viewing direction as

$$E(\lambda, \mathbf{x}) = e(\lambda, \mathbf{x}) \left((1 - \rho_f(\mathbf{x}))^2 R_\infty(\lambda, \mathbf{x}) + \rho_f(\mathbf{x}) \right) \quad (1)$$

where $\mathbf{x}$ denotes the spatial location on the image plane, λ the wavelength of the light, $e(\lambda, \mathbf{x})$ the spectrum of the light source, R_∞ the material reflectivity and ρ_f the Fresnel reflectance coefficient. Partial derivatives of E with respect to x and λ are denoted by subscripts E_x and E_λ, respectively.

A color invariant representation does not rely on accidental scene properties such as lighting and viewing direction and depends only on the material property R_∞. By exploring simplifying assumptions in Eq. (1), we can derive various invariant representations, as summarized in Table 1. The derived invariants E, W, C, N and H represent edge detectors that are invariant to various combinations of illumination changes, including scene geometry (i.e. does not detect shadow and shading edges), Fresnel reflections, and the intensity and color of the illuminant. For the complete derivations of the color invariants in Table 1, we refer to Section 1 of the supplementary material.

The Gaussian color model [24] is used to estimate E, E_λ and $E_{\lambda\lambda}$ from the RGB camera responses as

$$\begin{bmatrix} E(x,y) \\ E_\lambda(x,y) \\ E_{\lambda\lambda}(x,y) \end{bmatrix} = \begin{bmatrix} 0.06 & 0.63 & 0.27 \\ 0.3 & 0.04 & -0.35 \\ 0.34 & -0.6 & 0.17 \end{bmatrix} \begin{bmatrix} R(x,y) \\ G(x,y) \\ B(x,y) \end{bmatrix} \quad (2)$$

where x, y are pixel location in the image. Spatial derivatives E_x and E_y are calculated by convolving E with a Gaussian derivative kernel g with standard deviation σ, i.e.

$$E_x(x, y, \sigma) = \sum_{t \in \mathbb{Z}} E(t, y) \frac{\partial g(x - t, \sigma)}{\partial x} \quad (3)$$
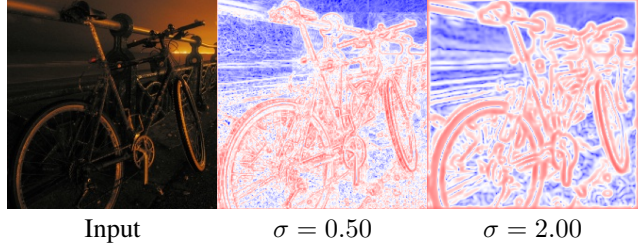

Figure 2: Color invariant representation W of the input image for two different values of σ. Note the trade-off between detail (small σ) and noise robustness (large σ).

and similarly for E_y, $E_{\lambda x}$, $E_{\lambda\lambda x}$, $E_{\lambda y}$ and $E_{\lambda\lambda y}$. Finally, the color invariant edge map is defined as the gradient magnitude of all relevant spatial derivatives as shown in Table 1.

The σ parameter in Eq. (3) determines the scale at which the image is convolved with the Gaussian derivative filters and as such the amount of detail preserved in the color invariant representation of an image. A small σ results in a detailed edge map but is more sensitive to noise, whereas a large σ is more robust but may omit important details. A visualization is given in Fig. 2 for color invariant W. Rather than fixing σ a-priori we implement the edge detector as a trainable layer to learn the task-specific optimal scale. The resulting Color Invariant Convolution (CIConv) is used as the input layer of the CNN and outputs a single channel representation onto which subsequent convolutional layers can be stacked. For computational simplicity we omit the square root from the gradient magnitude of the color invariants, and apply a log transformation and sample-wise normalization such that the distribution of the edge maps is close to standard normal. Furthermore, instead of directly optimizing σ, we train a scale parameter s such that $\sigma = 2^s$. This stabilizes training by reducing the backpropagation gradient for small values of s and ensures that σ is always positive. CIConv is thus defined as

$$\text{CIConv}(x, y) = \frac{\log\left(\text{CI}^2(x, y, \sigma = 2^s) + \epsilon\right) - \mu_\mathcal{S}}{\sigma_\mathcal{S}} \quad (4)$$

with CI the color invariant of choice from Tab. 1, $\mu_\mathcal{S}$ and $\sigma_\mathcal{S}$ the sample mean and standard deviation over $\log\left(\text{CI}^2 + \epsilon\right)$, and ϵ a small term added for numerical stability.

4. Experiments

4.1. Illumination robustness of CNNs

We investigate to what degree CIConv improves a CNN's robustness to accidental recording conditions by performing a classification experiment on a synthetic image dataset where we have accurate control over the illumination of the scene. The images are rendered from a subset

Invariant	Definition	SG	FR	II	IC
E	$E = \sqrt{E_x^2 + E_{\lambda x}^2 + E_{\lambda\lambda x}^2 + E_y^2 + E_{\lambda y}^2 + E_{\lambda\lambda y}^2}$	✗	✗	✗	✗
W	$W = \sqrt{W_x^2 + W_{\lambda x}^2 + W_{\lambda\lambda x}^2 + W_y^2 + W_{\lambda y}^2 + W_{\lambda\lambda y}^2}$, $W_x = \frac{E_x}{E}, \quad W_{\lambda x} = \frac{E_{\lambda x}}{E}, \quad W_{\lambda\lambda x} = \frac{E_{\lambda\lambda x}}{E}$	✗	✗	✓	✗
C	$C = \sqrt{C_{\lambda x}^2 + C_{\lambda\lambda x}^2 + C_{\lambda y}^2 + C_{\lambda\lambda y}^2}$, $C_{\lambda x} = \frac{E_{\lambda x}E - E_\lambda E_x}{E^2}, \quad C_{\lambda\lambda x} = \frac{E_{\lambda\lambda x}E - E_{\lambda\lambda}E_x}{E^2}$	✓	✗	✓	✗
N	$N = \sqrt{N_{\lambda x}^2 + N_{\lambda\lambda x}^2 + N_{\lambda y}^2 + N_{\lambda\lambda y}^2}$, $N_{\lambda x} = \frac{E_{\lambda x}E - E_\lambda E_x}{E^2}, N_{\lambda\lambda x} = \frac{E_{\lambda\lambda x}E^2 - E_{\lambda\lambda}E_x E - 2E_{\lambda x}E_\lambda E + 2E_\lambda^2 E_x}{E^3}$	✓	✗	✓	✓
H	$H = \sqrt{H_x^2 + H_y^2}, \quad H_x = \frac{E_{\lambda\lambda}E_{\lambda x} - E_\lambda E_{\lambda\lambda x}}{E_\lambda^2 + E_{\lambda\lambda}^2}$	✓	✓	✓	✗

Table 1: Overview of color invariant edge detectors [24] and their invariance properties to **S**cene **G**eometry, **F**resnel **R**eflections, **I**llumination **I**ntensity, **I**llumination **C**olor. E is a baseline intensity edge detector and is not invariant to any changes. Subscripts denote partial derivatives, where λ is the spectral derivative and x the spatial derivative of Eq. (1). Spatial derivatives for the y direction follow directly from the ones given for the x direction.

of the ShapeNet [12] dataset using the physically based renderer Mitsuba [31]. The scene is illuminated by a point light modeled as a black-body radiator with temperatures ranging between $[1900, 20000]K$ and an ambient light source. The training set contains 1,000 samples for each of the 10 object classes recorded under "normal" lighting conditions ($T = 6500K$). Multiple test sets with 300 samples per class are rendered for a variety of light source intensities and colors. Fig. 3 shows an overview of the illumination conditions represented in the test set.

CIConv improves illumination robustness We train a baseline ResNet-18 [26] and five models with the CIConv layer with invariants E, W, C, N and H, respectively. Training is done for 175 epochs with a batch size of 64 using SGD with momentum 0.9, weight decay 1e-4 and an initial learning rate of 0.05 with stepwise reduction by factor 0.1, step size 50. Data augmentation is performed in the form of random horizontal flips, random cropping and random rotations. The models are evaluated on both test sets and the average classification accuracy over three runs is shown in Fig. 4. The accuracy of the baseline RGB model quickly drops as lighting conditions start to diverge from the training set. The performance of the color invariant networks remains more stable with W consistently outperforming all others.

CIConv reduces feature map distribution shift The robustness of the color invariant networks compared to the baseline can be explained by analyzing the feature map activations of the networks. We calculate the mean feature map activation in different layers of the networks, averaged

over all samples in the Normal and Dark test sets. The histograms in Fig. 1 show that the intensity change between the normal and low light test sets caused a clear distribution shift throughout all network layers of the baseline model. In contrast, the CIConv layer with invariant W produces a domain invariant feature representation and consequently the distributions in the network are more aligned between the two domains. We quantify the distribution shift as the L2 distance between feature maps for the two domains, where again W yields the smallest distance. The L2 distances as well as histograms of the distributions of feature map activations for other color invariants are provided in section 2 of the supplementary material.

4.2. Day-night natural image classification

To verify that the properties of the color invariants also generalize to natural images we perform a classification experiment on a novel day-to-night dataset. We present the Common Objects Day and Night (CODaN) dataset, consisting of images from 10 common object classes recorded in both day and nighttime. It contains a daytime training set of 1,000 samples per class, a daytime validation set of 50 samples per class, and separate day and night test sets of 300 samples per class. CODaN is composed from the ImageNet [17], COCO [42] and ExDark [44] datasets. Samples of the day and night test sets are shown in Fig. 5.

Performance on natural images We trained color invariant versions of ResNet-18 on CODaN using the same settings as in 4.1, but without random cropping and with random brightness, contrast, hue and saturation augmentations. Table 2 shows the accuracy of the baseline and the color in-

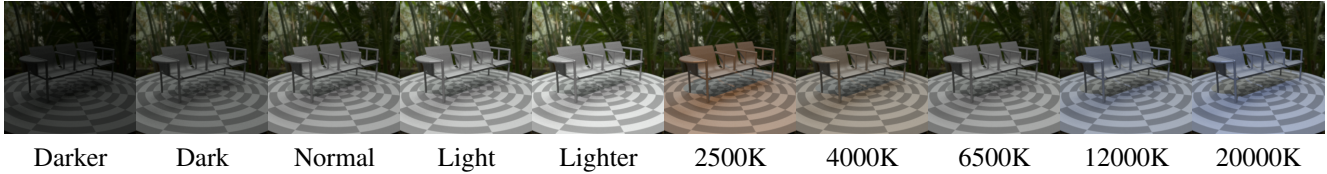Darker | Dark | Normal | Light | Lighter | 2500K | 4000K | 6500K | 12000K | 20000K

Figure 3: Sample from the synthetic classification dataset rendered from ShapeNet [12], shown in all illumination conditions represented in the test set. The five leftmost samples correspond to a varying light source intensity, whereas in the five rightmost samples a range of light source temperatures is shown. "Normal" and "6500K" are equivalent.
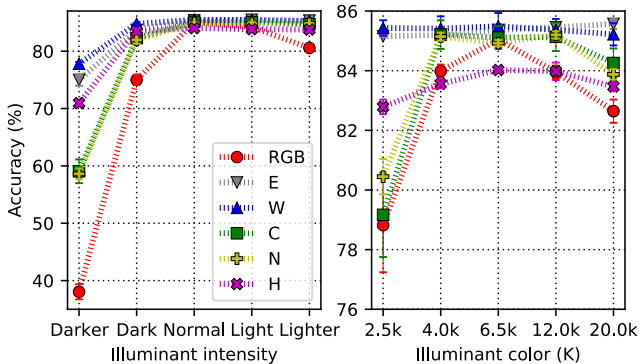

Figure 4: Classification accuracy of ResNet-18 with various color invariants on the synthetic ShapeNet dataset. RGB (not invariant) performance degrades when illumination conditions differ between train and test set, while color invariants remain more stable. W performs best overall.

Method	Day	Night
Baseline	80.39 ± 0.38	48.31 ± 1.33
E	79.79 ± 0.40	49.95 ± 1.60
W	$\mathbf{81.49 \pm 0.49}$	$\mathbf{59.67 \pm 0.93}$
C	78.04 ± 1.08	53.44 ± 1.28
N	77.44 ± 0.00	52.03 ± 0.27
H	75.20 ± 0.56	50.52 ± 1.34
Luminance	80.67 ± 0.32	51.37 ± 0.58
Normalized RGB	63.44 ± 1.52	41.66 ± 1.56
Comprehensive norm. [22]	70.52 ± 1.10	44.34 ± 1.57
Alvarez and Lopez [6]	64.41 ± 0.74	30.06 ± 0.57
Maddern et al. [46]	60.83 ± 0.98	33.04 ± 1.28
AdaBN [39]	79.72 ± 0.59	55.55 ± 1.07
Ablations	**Day**	**Night**
Baseline + norm.	63.43 ± 1.32	42.15 ± 0.98
Baseline + log + norm.	63.49 ± 0.55	41.90 ± 0.69
Baseline w/o color aug.	78.99 ± 0.59	36.00 ± 0.59
W w/o color aug.	79.71 ± 0.57	53.62 ± 0.88

Table 2: CODaN classification accuracy of a ResNet-18 architecture with various color invariants (top). W performs best. Ablation studies (bottom) show the individual effect of normalization, log scaling and photometric augmentations.

variant networks, averaged over three runs. Additionally, other color invariants (luminance, normalized RGB, comprehensive normalization [22] and others [6, 46]) are evaluated, which are implemented as a preprocessing step. We also consider a slightly adjusted version of AdaBN as a possible zero-shot domain adaptation method, which provides a significant performance increase by sampling the batch statistics for the Batch Normalization layers during test time for each individual batch. This is opposed to the original AdaBN method, where the batch statistics are calculated from the target domain dataset a priori. W outperforms all other models on the nighttime test set by a large margin. The luminance baseline performs surprisingly well, whereas the other non-trainable color invariants even result in a performance drop.

Color invariant transformations on natural images We visualize the E, W, C, N and H color invariant transformations of a day and night test sample (RGB) in Fig. 6. E being a non-invariant edge detector has low edge strengths in low intensity parts of the dark image. W on the other hand normalizes for intensity, yielding a more constant edge map. C, N and H are invariant to changes in scene geometry and

therefore do not detect edges with low color saturation, resulting in significant information loss. In addition, these invariants seem to be more amplifying the noise in low intensity parts of the image. Overall, W is able to 1) detect low intensity and low saturation edges and 2) suppress noise in low-intensity parts of the image, and therefore produces the most robust and informative edge map.

Learned vs. fixed scale We verify that CIConv learns the optimal scale by training the model with a range of fixed σ values, using invariant W. Fig. 7 shows the average accuracy over five runs. We observe that selecting the wrong scale σ has a detrimental effect on accuracy. When the scale is learnable, it converges to the optimal value for the daytime dataset, as indicated by the red cross in the figure. This value proves also optimal for the nighttime domain.

Figure 5: Samples from the day (source domain) and night (target domain) test sets of the CODaN dataset.
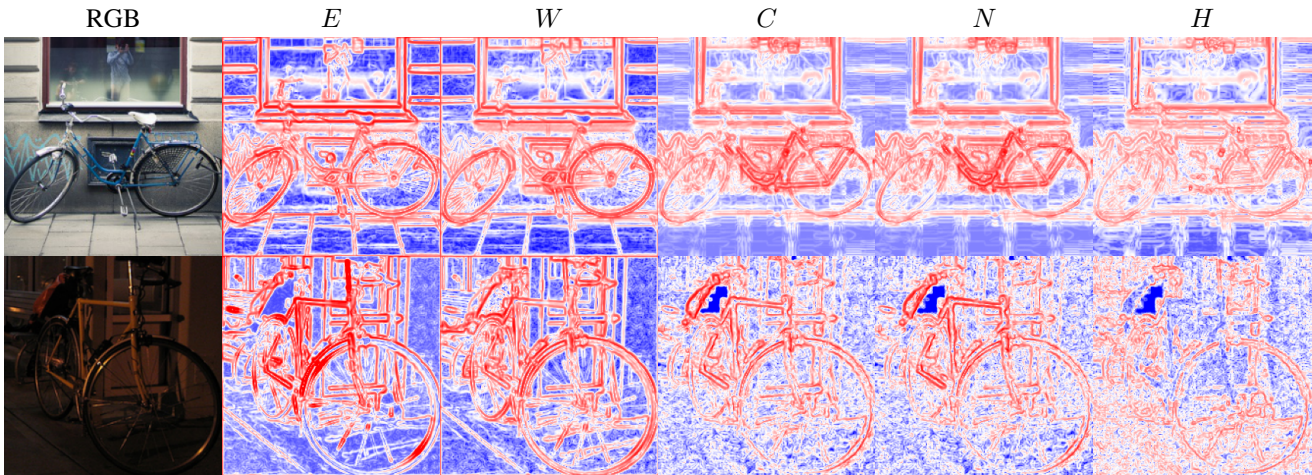

Figure 6: Color invariant visualizations of day and night samples from CODaN (red: positive; blue: negative values). E does not detect low intensity edges, whereas C, N and H do not detect edges that have low color saturation. W produces the most robust and informative edge map.
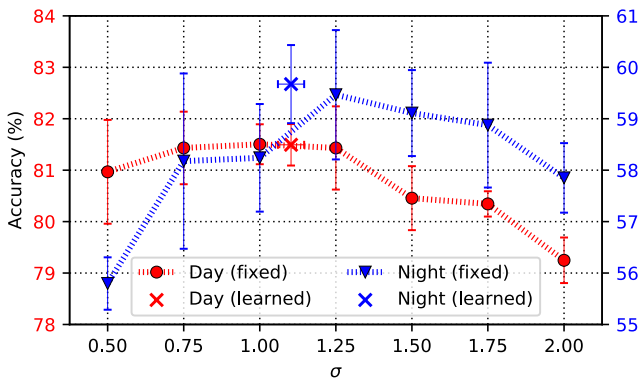

Figure 7: Performance on CODaN day (left y-axis) and night (right y-axis) test sets for various fixed values of σ. Learned σ and corresponding accuracies are indicated by crosses. CIConv learns the optimal value.

Ablation studies We evaluate whether simple log scaling and sample-wise normalization of RGB images, without applying a color invariant transformation, can achieve the

same improved performance on the nighttime test set. Furthermore, we investigate how the baseline and W networks perform when trained without brightness, contrast, hue and saturation augmentations. The results are shown in the bottom part of Table 2. Normalization, both with and without log scaling, does not yield better performance for the baseline model. This indicates that addressing the distribution shift between the source and target domain observed in the feature map activations of a network requires more than simple intensity normalization of the input sample. Moreover, photometric augmentations mostly seem to benefit the baseline network, whereas the model with color invariant W is inherently more robust to illumination changes. Both results underscore the importance and effectiveness of the color invariant transformation.

4.3. Semantic segmentation

We perform a semantic segmentation experiment using the RefineNet [41] architecture with ResNet-101 and W-ResNet-101 feature extractors pre-trained on the ImageNet [17] dataset. The segmentation model is trained on

Method	Nighttime Driving	Dark Zurich
Trained on source data only		
RefineNet [41]	34.1	30.6
W-RefineNet [ours]	**41.6**	**34.5**
RefineNet-AdaBN [39]	36.3	31.3
Trained on source and target data		
ADVENT [74]	34.7	29.7
BDL [40]	34.7	30.8
AdaptSegNet [64]	34.5	30.4
DMAda [16]	41.6	32.1
Day2Night [58]	45.1	-
GCMA [56]	45.6	42.0
MGCDA [57]	**49.4**	**42.5**

Table 3: Segmentation performance on Nighttime Driving [16] and Dark Zurich [56], reported as mIoU scores. *W*-RefineNet outperforms other methods trained only on daytime data and has competitive performance to methods also using nighttime images.

Method	Tokyo 24/7 (mAP)
Trained on source data only	
VGG GeM [52]	79.4
W-VGG GeM [ours]	83.3
ResNet101 GeM [52]	85.0
W-ResNet101 GeM [ours]	**88.3**
EdgeMAC [53]	75.9
U-Net jointly [32]	79.8
CLAHE [85]	84.1
EdgeMAC + VGG GeM [32]	85.4
Trained on source and target data	
VGG GeM [52]	79.8
U-Net jointly [32]	86.5
CLAHE [85]	87.0
EdgeMAC + CLAHE [32]	**90.5**
EdgeMAC + U-Net jointly [32]	90.0

Table 4: Place recognition results on the Tokyo 24/7 dataset [62]. VGG GeM with our CIConv layer outperforms all other methods trained on daytime data. + denotes an ensemble of different models.

the training set of the CityScapes [14] dataset containing 2,975 densely annotated daytime street images and evaluated on the 50 coarsely annotated street images from Nighttime Driving [16] and the 151 densely annotated images from the Dark Zurich [56] test set. We perform training using SGD with momentum 0.9, weight decay 1e-4 and an initial learning rate of 0.1 which is step-wise reduced by a factor 0.1 after every 30 epochs. All input images are resized to 1024x512 pixels and randomly cropped to 768x384 pixels, allowing a batch size of 6 on 2 GeForce GTX 1080 Ti GPUs. Data augmentation is applied by random scaling, brightness-, contrast- and hue-shifting, and horizontal flipping. Inference is done on 1024x512 samples without cropping.

Results are shown in Table 3 as the mean Intersection-over-Union (mIoU). Results for other methods are taken from their corresponding papers. The color invariant *W*-RefineNet significantly outperforms the vanilla RefineNet and RefineNet-AdaBN models, which are also trained only on source domain data, and has competitive performance compared to methods trained on both source and target domain data. Qualitative segmentation results are shown in Fig. 8. Detailed per-class scores are included in section 4 of the supplementary material.

4.4. Visual place recognition (VPR)

We present results for VPR task in two phases: first, we compare against a similar work for place recognition based on a learnable normalisation of images [32], and then we

benchmark place representations based on color-invariant trained CNNs on an additional dataset, evaluation metric, and descriptor type to show broader applicability within VPR.

Learnable normalisation. We use the Tokyo 24/7 day-night place recognition dataset [62] for this purpose, and follow the evaluation procedure described in [32]. To obtain place representations, the VGG Generalized Mean Pooling (GeM) [52] network is prepended with our CIConv layer (*W*-VGG GeM) and trained on the Retrieval-SfM dataset as described in [52]. The train dataset contains query images as well as both positive and negative target images of places photographed in daytime conditions. The results are reported as the mean Average Precision (mAP) in Table 4. Results of competing methods are borrowed from Tables 1 and 2 in [32]. It can be observed that our method outperforms all models trained on daytime data only and achieves competitive results to the current state-of-the-art, which is an ensemble of two models trained on both daytime and nighttime data.

Broader VPR applicability. Here, we use the two outdoor day-night datasets from VPRBench [81]: Gardens Point and Tokyo 24/7, where latter's evaluation is similar to the previous experiment but using Recall@1 as the evaluation metric in this case for both the datasets. For the Gardens Point dataset, we consider two settings: A (Appearance only) with only day-night variations and more challenging A+V (Appearance + Viewpoint) where viewpoint is also laterally shifted. We consider three descriptor pool-

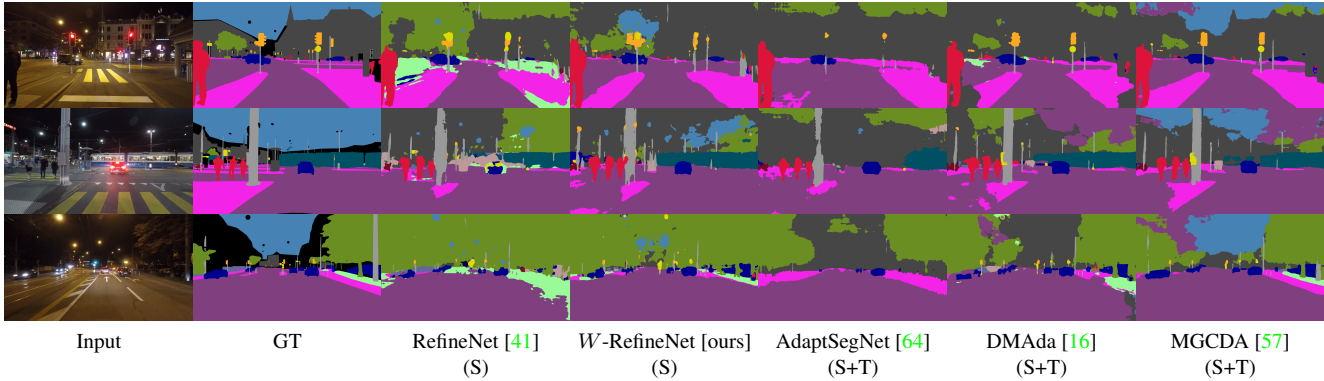| Input | GT | RefineNet [41] (S) | W-RefineNet [ours] (S) | AdaptSegNet [64] (S+T) | DMAda [16] (S+T) | MGCDA [57] (S+T) |

Figure 8: Qualitative semantic segmentation results on the Dark Zurich [56] dataset. S and T indicate whether the model was trained on the source or target domain, respectively.

Method	GP:A+V	GP:A	Tokyo 24/7
AP-GeM [54]	0.87	0.92	0.91
DenseVLAD [63]	0.81	0.89	0.89
R101 MAC [60]	0.51	0.56	0.20
R101 Flat [59]	0.56	0.68	0.84
R101 GeM [52]	0.90	0.96	0.91
W-R101 MAC [ours]	0.53	0.70	0.20
W-R101 Flat [ours]	0.61	0.91	0.85
W-R101 GeM [ours]	**0.94**	**0.97**	**0.93**

Table 5: Recall@1 for VPR using different feature pooling types on Gardens Point (GP) and Tokyo 24/7 dataset. Color-invariant layer (W) based networks outperform their vanilla counterparts with W-R101-GeM achieving state-of-the-art results.

ing types here using ImageNet-trained ResNet-101 (R101) as the backbone network: Maximum Activations of Convolutions (MAC) [60], flattened tensor (Flat) [59] and GeM, where only GeM is further trained on image retrieval task as described in the previous subsection. For all three descriptor types, we compute results for training with and without the prepended color invariant layer. Additionally, we compare against state-of-the-art VPR methods: DenseVLAD [63] and AP-GeM [54].

In Table 5, it can be observed that W-R101 GeM achieves state-of-the-art results for all datasets. Furthermore, all methods based on color invariant perform better than their vanilla counterparts, including the Flat and MAC descriptors. This shows that color invariant networks provide robust place representation for different pooling types even without VPR-specific training.

5. Discussion

The image formation model that lies at the foundation of the color invariants used in the CIConv layer is based on certain simplifying assumptions, such as purely matte reflections, non-transparent materials and a single, spatially uniform light source. Although most natural scenes do not satisfy these strict conditions, our results show that CNNs nevertheless do benefit from prior information derived from such approximate models. Moreover, current publicly available datasets, including the ones used in our experiments, are not appropriate for physics-based vision due to various artifacts introduced in post-processing steps (see Discussion in [47]). CIConv and other physics based methods can therefore only reach their full potential when sufficient attention is paid to preserving the physical correctness of the data during image capturing.

The robustness of color invariants to illumination changes comes at the loss of some discriminative power [24]. The CIConv layer transforms the input image into an edge map representation that is no longer sensitive to the intensity and color of the light source, but as a side effect also removes valuable color information. We found that naively concatenating color invariants with the RGB input degrades performance, see section 3 of the supplementary material. Future research should therefore focus on implementing an adaptive mechanism for optimally combining color information and color invariant edge information.

Zero-shot domain adaptation is a promising method for reducing the data dependency and the corresponding data collection and annotation costs in computer vision. We therefore hope that this paper inspires future research on integrating physics priors into neural networks.

Acknowledgements

References

[1] Mahmoud Afifi and Michael Scott Brown. What else can fool deep learning? addressing color constancy errors on deep neural network performance. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 243–252, 2019. 1, 3

[2] Zeynep Akata, Florent Perronnin, Zaid Harchaoui, and Cordelia Schmid. Label-embedding for image classification. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1425–1438, 2015. 2

[3] N. Alshammari, S. Akcay, and T. P. Breckon. On the impact of illumination-invariant image pre-transformation for contemporary automotive semantic scene understanding. In *2018 IEEE Intelligent Vehicles Symposium (IV)*, pages 1027–1032, 2018. 3

[4] Naif Alshammari, Samet Akçay, and T. Breckon. Multi-task learning for automotive foggy scene understanding via domain adaptation to an illumination-invariant representation. *ArXiv*, abs/1909.07697, 2019. 3

[5] J. A. Escobedo Alvarez, Antonio Lopez, and Ramón Baldrich. Illuminant-invariant model-based road segmentation. *2008 IEEE Intelligent Vehicles Symposium*, pages 1175–1180, 2008. 3

[6] J. M. A. Alvarez and A. M. Lopez. Road detection based on illuminant invariance. *IEEE Transactions on Intelligent Transportation Systems*, 12(1):184–193, March 2011. 2, 3, 5

[7] Asha Anoosheh, Torsten Sattler, Radu Timofte, Marc Pollefeys, and Luc Van Gool. Night-to-day image translation for retrieval-based localization. *2019 International Conference on Robotics and Automation (ICRA)*, pages 5958–5964, 2019. 2

[8] Kobus Barnard, Graham Finlayson, and Brian Funt. Color constancy for scenes with varying illumination. *Computer vision and image understanding*, 65(2):311–321, 1997. 2

[9] A. S. Baslamisli, T. T. Groenestege, P. Das, H. A. Le, S. Karaoglu, and T. Gevers. Joint learning of intrinsic images and semantic segmentation. In *European Conference on Computer Vision*, 2018. 3

[10] Anil S. Baslamisli, Hoang-An Le, and Theo Gevers. CNN based learning using reflection and retinex models for intrinsic image decomposition. *CoRR*, abs/1712.01056, 2017. 2

[11] Gertjan J Burghouts and Jan-Mark Geusebroek. Performance evaluation of local colour invariants. *Computer Vision and Image Understanding*, 113(1):48–62, 2009. 2

[12] Angel X. Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, Jianxiong Xiao, Li Yi, and Fisher Yu. ShapeNet: An Information-Rich 3D Model Repository. Technical Report arXiv:1512.03012 [cs.GR], Stanford University — Princeton University — Toyota Technological Institute at Chicago, 2015. 4, 5

[13] S. W. Cho, N. R. Baek, J. H. Koo, M. Arsalan, and K. R. Park. Semantic segmentation with low light images by modified cyclegan-based image enhancement. *IEEE Access*, 8:93561–93585, 2020. 2

[14] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 7

[15] Peter I. Corke, Rohan Paul, Winston Churchill, and Paul Newman. Dealing with shadows: Capturing intrinsic scene appearance for image-based outdoor localisation. *2013 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 2085–2092, 2013. 3

[16] D. Dai and L. V. Gool. Dark model adaptation: Semantic image segmentation from daytime to nighttime. In *ITSC*, pages 3819–3824, Nov 2018. 1, 2, 7, 8

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*, 2009. 4, 6

[18] S. Di, Q. Feng, C. Li, M. Zhang, H. Zhang, S. Elezovikj, C. C. Tan, and H. Ling. Rainy night scene understanding with near scene semantic adaptation. *IEEE Transactions on Intelligent Transportation Systems*, pages 1–9, 2020. 2

[19] Graham D. Finlayson, Mark S. Drew, and Cheng Lu. Entropy minimization for shadow removal. *International Journal of Computer Vision*, 85:35–57, 2009. 3

[20] Graham D. Finlayson and Steven D. Hordley. Color constancy at a pixel. *Journal of the Optical Society of America. A, Optics, image science, and vision*, 18 2:253–64, 2001. 2

[21] 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*, 28:59–68, 2006. 3

[22] Graham D. Finlayson, Bernt Schiele, and James L. Crowley. Comprehensive colour image normalization. In Hans Burkhardt and Bernd Neumann, editors, *Computer Vision — ECCV'98*, pages 475–490, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg. 5

[23] Brian V. Funt and Graham D. Finlayson. Color constant color indexing. *IEEE transactions on Pattern analysis and Machine Intelligence*, 17(5):522–529, 1995. 2

[24] J. M. Geusebroek, R. van den Boomgaard, A. W. M. Smeulders, and H. Geerts. Color invariance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(12):1338–1350, 2001. 2, 3, 4, 8

[25] Theo Gevers and Arnold WM Smeulders. Color-based object recognition. *Pattern recognition*, 32(3):453–464, 1999. 2

[26] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 4

[27] Judy Hoffman, Eric Tzeng, Taesung Park, Jun-Yan Zhu, Phillip Isola, Kate Saenko, Alexei Efros, and Trevor Darrell. CyCADA: Cycle-consistent adversarial domain adaptation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1989–1998, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR. 2

[28] Judy Hoffman, Dequan Wang, Fisher Yu, and Trevor Darrell. Fcns in the wild: Pixel-level adversarial and constraint-based adaptation. *CoRR*, abs/1612.02649, 2016. 2

[29] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *CoRR*, abs/1502.03167, 2015. 1, 2

[30] Masato Ishii, Takashi Takenouchi, and Masashi Sugiyama. Zero-shot domain adaptation based on attribute information. In *ACML*, 2019. 2

[31] Wenzel Jakob. Mitsuba renderer, 2010. http://www.mitsuba-renderer.org. 4

[32] Tomás Jenícek and Ondřej Chum. No fear of the dark: Image retrieval under varying illumination conditions. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9695–9703, 2019. 2, 7

[33] O. Kayhan and J.C. van Gemert. On translation invariance in CNNs: Convolutional layers can exploit absolute spatial location. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[34] Taeyoung Kim, Yu-Wing Tai, and Sung eui Yoon. Pca based computation of illumination-invariant space for road detection. *2017 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 632–640, 2017. 3

[35] T. Krajník, J. Blažíček, and J. M. Santos. Visual road following using intrinsic images. In *2015 European Conference on Mobile Robots (ECMR)*, pages 1–6, 2015. 3

[36] P. Kubelka and F. Munk. Ein beitrag zur optik der farbanstriche. In *Zeitung fur Technische Physik*, volume 12, page 593, 1999. 2, 3

[37] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Attribute-based classification for zero-shot visual object categorization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 36(3):453–465, 2013. 2

[38] Ruoteng Li, Loong Fah Cheong, and Robby T. Tan. Heavy rain image restoration: Integrating physics model and conditional adversarial learning. *CoRR*, abs/1904.05050, 2019. 2

[39] Yanghao Li, Naiyan Wang, Jianping Shi, Jiaying Liu, and Xiaodi Hou. Revisiting batch normalization for practical domain adaptation. *CoRR*, abs/1603.04779, 2016. 1, 2, 5, 7

[40] Yunsheng Li, Lu Yuan, and Nuno Vasconcelos. Bidirectional learning for domain adaptation of semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019. 7

[41] G. Lin, A. Milan, C. Shen, and I. Reid. RefineNet: Multi-path refinement networks for high-resolution semantic segmentation. In *CVPR*, July 2017. 6, 7, 8

[42] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. 4

[43] Yancong Lin, Silvia L Pintea, and Jan C van Gemert. Deep hough-transform line priors. In *European Conference on Computer Vision*, pages 323–340. Springer, 2020. 2

[44] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer Vision and Image Understanding*, 178:30–42, 2019. 4

[45] Mingsheng Long and Jianmin Wang. Learning transferable features with deep adaptation networks. *CoRR*, abs/1502.02791, 2015. 2

[46] 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, May 2014. 2, 3, 5

[47] Bruce A. Maxwell, Casey A. Smith, Maan Qraitem, Ross Messing, Spencer Whitt, Nicolas Thien, and Richard M. Friedhoff. Real-time physics-based removal of shadows and shading from road surfaces. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1277–1285, 2019. 3, 8

[48] Thomas Mensink, Efstratios Gavves, and Cees GM Snoek. Costa: Co-occurrence statistics for zero-shot classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2441–2448, 2014. 2

[49] Mohammad Norouzi, Tomas Mikolov, Samy Bengio, Yoram Singer, Jonathon Shlens, Andrea Frome, Greg S Corrado, and Jeffrey Dean. Zero-shot learning by convex combination of semantic embeddings. *arXiv preprint arXiv:1312.5650*, 2013. 2

[50] Kuan-Chuan Peng, Ziyan Wu, and Jan Ernst. Zero-shot deep domain adaptation. In *ECCV*, 2017. 2

[51] H. Porav, W. Maddern, and P. Newman. Adversarial training for adverse conditions: Robust metric localisation using appearance transfer. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1011–1018, 2018. 2

[52] Filip Radenović, Giorgos Tolias, and Ondřej Chum. Fine-tuning cnn image retrieval with no human annotation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41:1655–1668, 2017. 7, 8

[53] F. Radenović, G. Tolias, and O. Chum. Deep shape matching. *ECCV*, 2018. 7

[54] Jerome Revaud, Jon Almazán, Rafael S Rezende, and Cesar Roberto de Souza. Learning with average precision: Training image retrieval with a listwise loss. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 5107–5116, 2019. 8

[55] Eduardo Romera, Luis Miguel Bergasa, Kailun Yang, Jose M. Álvarez, and Rafael Barea. Bridging the day and night domain gap for semantic segmentation. *2019 IEEE Intelligent Vehicles Symposium (IV)*, pages 1312–1318, 2019. 1, 2

[56] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Guided curriculum model adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 2, 7, 8

[57] Christos Sakaridis, Dengxin Dai, and Luc Van Gool. Map-guided curriculum domain adaptation and uncertainty-aware evaluation for semantic nighttime image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 1, 2, 7, 8

[58] Lei Sun, Kaiwei Wang, Kailun Yang, and Kaite Xiang. See clearer at night: towards robust nighttime semantic segmentation through day-night image conversion. In Judith Dijk, editor, *Artificial Intelligence and Machine Learning in Defense Applications*, page 8, Strasbourg, France, Sept. 2019. SPIE. 2, 7

[59] Niko Sünderhauf, Sareh Shirazi, Feras Dayoub, Ben Upcroft, and Michael Milford. On the performance of convnet features for place recognition. In *2015 IEEE/RSJ international conference on intelligent robots and systems (IROS)*, pages 4297–4304. IEEE, 2015. 8

[60] Giorgos Tolias, Ronan Sicre, and Hervé Jégou. Particular object retrieval with integral max-pooling of cnn activations. *arXiv preprint arXiv:1511.05879*, 2015. 8

[61] Nergis Tomen and Jan van Gemert. Spectral leakage and rethinking the kernel size in cnns. *arXiv preprint arXiv:2101.10143*, 2021. 2

[62] A. Torii, R. Arandjelović, J. Sivic, M. Okutomi, and T. Pajdla. 24/7 place recognition by view synthesis. In *CVPR*, 2015. 7

[63] Akihiko Torii, Relja Arandjelovic, Josef Sivic, Masatoshi Okutomi, and Tomas Pajdla. 24/7 place recognition by view synthesis. In *IEEE Conf. Comput. Vis. Pattern Recog.*, pages 1808–1817, 2015. 8

[64] Y. Tsai, W. Hung, S. Schulter, K. Sohn, M. Yang, and M. Chandraker. Learning to adapt structured output space for semantic segmentation. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7472–7481, 2018. 7, 8

[65] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. *CoRR*, abs/1702.05464, 2017. 2

[66] Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *CoRR*, abs/1412.3474, 2014. 2

[67] Ben Upcroft, Colin McManus, Winston Churchill, William P. Maddern, and Paul Newman. Lighting invariant urban street classification. *2014 IEEE International Conference on Robotics and Automation (ICRA)*, pages 1712–1718, 2014. 3

[68] Gregor Urban, Krzysztof J Geras, Samira Ebrahimi Kahou, Ozlem Aslan, Shengjie Wang, Abdelrahman Mohamed, Matthai Philipose, Matt Richardson, and Rich Caruana. Do deep convolutional nets really need to be deep and convolutional? In *ICLR*, 2016. 2

[69] A. Valada, J. Vertens, A. Dhall, and W. Burgard. Adapnet: Adaptive semantic segmentation in adverse environmental conditions. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4644–4651, 2017. 2

[70] Koen Van De Sande, Theo Gevers, and Cees Snoek. Evaluating color descriptors for object and scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1582–1596, 2009. 2

[71] Joost Van de Weijer, Theo Gevers, and Andrew D Bagdanov. Boosting color saliency in image feature detection. *IEEE transactions on pattern analysis and machine intelligence*, 28(1):150–156, 2005. 2

[72] Joost Van de Weijer, Theo Gevers, and J-M Geusebroek. Edge and corner detection by photometric quasi-invariants. *IEEE transactions on pattern analysis and machine intelligence*, 27(4):625–630, 2005. 2

[73] Johan Vertens, Jannik Zürn, and Wolfram Burgard. Heatnet: Bridging the day-night domain gap in semantic segmentation with thermal images, 2020. 2

[74] Tuan-Hung Vu, Himalaya Jain, Maxime Bucher, Mathieu Cord, and Patrick Pérez. Advent: Adversarial entropy minimization for domain adaptation in semantic segmentation. In *CVPR*, 2019. 7

[75] Jinghua Wang and Jianmin Jiang. Conditional coupled generative adversarial networks for zero-shot domain adaptation. In *ICCV*, October 2019. 2

[76] Mei Wang and Weihong Deng. Deep visual domain adaptation: A survey. *Neurocomputing*, 312:135–153, 2018. 1, 2

[77] Qian Wang, Penghui Bu, and Toby P. Breckon. Unifying unsupervised domain adaptation and zero-shot visual recognition. *CoRR*, abs/1903.10601, 2019. 2

[78] M. Wulfmeier, A. Bewley, and I. Posner. Addressing appearance change in outdoor robotics with adversarial domain adaptation. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 1551–1558, 2017. 1, 2

[79] Yongqin Xian, Christoph H Lampert, Bernt Schiele, and Zeynep Akata. Zero-shot learning-a comprehensive evaluation of the good, the bad and the ugly. *IEEE transactions on pattern analysis and machine intelligence*, 2018. 2

[80] Yongxin Yang and Timothy M. Hospedales. Zero-shot domain adaptation via kernel regression on the grassmannian. *CoRR*, abs/1507.07830, 2015. 2

[81] Mubariz Zaffar, Shoaib Ehsan, Michael Milford, David Flynn, and Klaus McDonald-Maier. Vpr-bench: An open-source visual place recognition evaluation framework with quantifiable viewpoint and appearance change. *arXiv preprint arXiv:2005.08135*, 2020. 7

[82] Richard Zhang. Making convolutional networks shift-invariant again. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, volume 97, pages 7324–7334, 2019. 2

[83] Ziming Zhang and Venkatesh Saligrama. Zero-shot learning via joint latent similarity embedding. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016. 2

[84] Yuan Zhou and Kangming Yan. Domain adaptive adversarial learning based on physics model feedback for underwater image enhancement. *ArXiv*, abs/2002.09315, 2020. 2

[85] Karel Zuiderveld. *Contrast Limited Adaptive Histogram Equalization*, page 474–485. Academic Press Professional, Inc., USA, 1994. 7