DB-GAN: Boosting Object Recognition Under Strong Lighting Conditions

Luca Minciullo *
Toyota Motor Europe
luca.minciullo@toyota-europe.com

Fabian Manhardt*
Technical University of Munich
fabian.manhardt@tum.de

Kei Yoshikawa
Toyota Motor Europe
kei.yoshikawa@toyota-europe.com

Sven Meier
Toyota Motor Europe
sven.meier@toyota-europe.com

Federico Tombari†
Technical University of Munich
tombari@in.tum.de

Norimasa Kobori
Woven CORE, Inc.
norimasa.kobori@tri-ad.global

Abstract

Driven by deep learning, object recognition has recently made a tremendous leap forward. Nonetheless, its accuracy often still suffers from several sources of variation that can be found in real-world images. Some of the most challenging variations are induced by changing lighting conditions. This paper presents a novel approach for tackling brightness variation in the domain of 2D object detection and 6D object pose estimation. Existing works aiming at improving robustness towards different lighting conditions are often grounded on classical computer vision contrast normalisation techniques or the acquisition of large amounts of annotated data in order to achieve invariance during training. While the former cannot generalise well to a wide range of illumination conditions, the latter is neither practical nor scalable. Hence, We propose the usage of Generative Adversarial Networks in order to learn how to normalise the illumination of an input image. Thereby, the generator is explicitly designed to normalise illumination in images so to enhance the object recognition performance. Extensive evaluations demonstrate that leveraging the generated data can significantly enhance the detection performance, outperforming all other state-of-the-art methods. We further constitute a natural extension focusing on white balance variations and introduce a new dataset for evaluation.

1. Introduction

Due to its wide range of applications, localising objects in natural images is one of the most studied fields in computer vision [41, 55, 26, 20, 52, 12, 13]. Recently, driven by deep learning and the accessibility of large-scale datasets such as ImageNet [6] or Open Images [22], there has been tremendous improvement in terms of detection accuracy as well as the number of objects that can be recognized simultaneously [25, 27, 36, 10, 20, 33, 24]

Despite the undeniable advances, several open challenges still remain to be solved. Some of the most prominent being robustness towards illumination [17, 35, 51], viewpoint changes [28], occlusion, as well as handling the

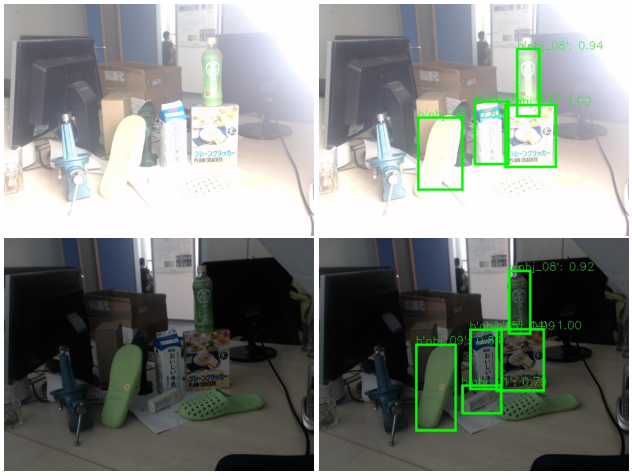

Figure 1. **Detection under strong lighting variations.** Although the input image is subject to strong light from the side, we can still detect almost all objects taken from Toyota Light [12] (top). Similarly, we are able to robustly detect these objects when little light is available (bottom). In each row the input image is shown on the left, while SSD detections are shown on the right.

*These authors contributed equally to this work
†Federico Tombari is now working at Google

synthetic-to-real domain gap [13, 46].

Real-world environments commonly possess a large variation of illumination conditions. For instance, applications involving outdoor scenes are often exposed to strong changes in illumination. In autonomous driving, cars oftentimes operate in extreme scenarios such as direct strong sunlight during the day or almost no light at night. Similarly, indoor vision systems often suffer from challenging lighting conditions. Noteworthy, nearby windows or inside refrigerators the contrast ratio be 1000:1 or higher. These challenges commonly go unnoticed when training on large scale datasets. However, many practical applications deal with objects categories or instances that are not part of benchmark datasets. Therefore, training data needs to be collected from scratch and the acquisition of data with the required variation is problematic. This is particularly true for 6D pose estimation, since annotating the 6D pose of an object is very difficult, time consuming, and error-prone [13]. For this reason, increasing the capability of models leveraging only synthetic data is of high interest [20, 46, 13, 44]. As a consequence, in this work we focus on robustness towards brightness and color with a particular focus on synthetic data.

In this paper we introduce our novel method to improve 2D object detection and 6D object pose estimation, which we call Detector-Boost GAN(DB-GAN) - a GAN-based architecture for illumination normalisation (*c.f*. Figure 1). Our method is essentially trained to perform illumination normalisation by means of generating images tailored to the capabilities of the object detector. By back-propagating the detection loss, DB-GAN learns to eradicate the weaknesses of the detector and strengthen its performance. Our method does not need prior information on the input image and is able to automatically recover normalised texture under dark, bright as well as non-uniform light conditions. DB-GAN is capable of outperforming all related state-of-the-art methods on two standard benchmark datasets, TUD Light and Toyota Light [12]. We also introduce a new dataset named Toyota TrueBlue, aimed at assessing robustness to white balance changes. Our approach is able to achieve significant mAP improvements on all datasets compared to our baseline detectors and other existing works. Noteworthy, despite focusing on improving detectors, our method can be potentially leveraged to enhance performance of various computer vision tasks.

In summary, we make the following contributions. i) We propose a novel architecture which learns to generate images in order to facilitate further detection under strong illumination changes. ii) We introduce Toyota TrueBlue, a new dataset focusing explicitly on robustness to change in white balance and iii) experimentally demonstrate that DB-GAN significantly enhances performance both in 2D and 3D, outperforming all related methods.

2. Related Work

In this section we provide an overview on previous works in illumination normalisation. Since we employ Generative Adversarial Networks (GANs) to normalize images, we also briefly outline the most important works in the GAN literature.

2.1. Generative Adversarial Networks

Generative Adversarial Networks(GANs) [8] are one of most important recent advances in generative models. GANs train in alternation two deep learning architectures: a generator and a discriminator. While the generator produces realistically looking images, the discriminator attempts to distinguish images coming from the generator from images sampled from the true distribution. The networks are trained jointly in a min-max game fashion, converging in an equilibrium in which the discriminator is not capable of distinguishing real from fake. Inspired by [8], Isola *et al*. employ Conditional GANs [30] for image translation between two domains [14]. Here the generated samples are also conditioned on the input sample, meaning that the discriminator always receives a pair of images. Accordingly, the discriminator is required to distinguish whether the generated output is consistent with the input and correctly translates to the target domain. Similarly, Cycle-GAN, proposed in [59] also carries out domain translation, but without the need for paired data. SINGAN [39] leverages a sequence of generators learning to reconstruct texture at different resolutions and can be trained using a single high resolution training image.

Some existing GAN based works have been introduced in the context of object detection. While, Wang *et al*. [49] leverage GANs for knowledge distillation, Bai *et al*. [2] focuses on improving the detection of small objects. In [58], the authors propose to use weakly supervised object discovery for the detection of vehicles in high resolution remote sensing images. Wang *et al*. [50] propose an adversarial mask generation approach to improve occlusion and deformation robustness in object detection. Finally, other works [7, 15] use GAN generators to produce instance level segmentation masks for either weakly supervised [7] or unpaired data based object detection [15].

Nevertheless, to the best of our knowledge, none of the mentioned works have been used to improve illumination robustness.

2.2. Illumination Normalisation

In this section we specifically cover illumination normalization, image enhancement and color constancy approaches with a special focus on GAN-based solutions.

Local Contrast Normalization [16], was introduced as a pre-processing step to mimic the behaviour of the V1 cells

in the cortical area of the brain. A few deep-learning approaches for robustness towards illumination changes have also been proposed. Krizhevsky *et al*. [21] introduce Local Response Normalization as a brightness normalisation module to be applied after non-linearities in deep architectures. Rad *et al*. [35] propose to learn the parameters of a generalization of the Difference-of-Gaussians(DoG) method using CNNs. Thereby, the DoG parameters are learned end-to-end with respect to object detection and 6D object pose estimation. Nonetheless, this method is inherently restricted by the capacity of DoGs for normalisation. Other works [23, 48] perform illumination estimation for modality fusion of thermal and color inputs [23] and image enhancement [48]. Several additional approaches have been introduced for general image enhancement [53, 54, 9, 31].

GAN-based approaches In [42], the authors leverage GANs (*i.e*. Angular-GAN) to remove light and shadows from RGB images. Their method is fully-supervised and uses synthetic training samples generated with GTA-V. Jiang *et al*. propose EnlightenGAN[17] for transforming dark into bright images and vice-versa. The architecture is inspired by Cycle-GAN[59], hence, eradicating the need for paired images during training. However, prior knowledge on whether the input image is *too dark* or *too bright* is required. Furthermore, this method assumes that the input image is acquired under uniform lighting, which is rarely the case in practical scenarios. Wei *et al*. recently introduce Retinex-Net[51], an end-to-end trainable architecture for low-light image enhancement. [51] decomposes the image into reflectance and illumination, prior to adjusting illumination. Nonetheless, they require paired low-light/normal-light data for training. Zhang *et al*. [57] propose a GAN base architecture to deal with illumination robustness in face recognition. They learn a illumination invariant latent space by means of adversarial training. Sakkos *et al*. [38] use two GAN generators to produce both low-light and bright images and then perform semantic segmentation on the difference image in a multi-task setting. Finally, Chen *et al*. [4] propose a GAN-based image enhancement approach.

3. Methodology

In this work, we propose a novel method for illumination normalisation in RGB images. The network is grounded on an Encoder-Decoder architecture, leveraging recent advances in GANs to further enhance the reconstruction quality. The core novelty of this work lies in the additional back-propagation of a detection loss, while training the GAN. This implicitly forces the network to generate images, which simplify latter object detection despite contrary conditions such as very strong illumination. Unlike previous works [17, 51] our method does not require prior knowledge of the input image as well as any real data for training.

In this section, we explain the technical details of our proposed method.

3.1. DB-GAN for Detection-Driven Reconstruction

Let $\mathcal{I}$ be any image space and $\bar{\mathcal{I}}$ be the subset of $\mathcal{I}$ whose elements possess uniform lighting. Assumed an acquired set of image pairs (I, $\bar{I}$), where $I \in \mathcal{I}$, and $\bar{I} \in \bar{\mathcal{I}}$, the illumination normalised version of $\mathcal{I}$. In addition, we assume that all objects of interest are annotated in the form of either bounding boxes or 6D poses. In the following sections we describe how we construct a dataset with these characteristics without any human labelling.

We want to learn a mapping from the domain $\mathcal{I}$ to the illumination-free domain $\bar{\mathcal{I}}$. To this end, we employ a GAN based architecture, following recent success of adversarial models at image generation tasks[59, 39, 14, 42]. To avoid losing details in the reconstructed image [14, 17, 3, 19], our generator $\mathcal{G}$ is based on an encoder-decoder architecture with skip connections [37]. Given an image pair (I, $\bar{I}$) the generator has to learn to normalise the input image according to

$$\hat{I} = \mathcal{G}(I). \tag{1}$$

Since we assume pairs of images, we can learn the mapping from $\mathcal{I}$ to $\bar{\mathcal{I}}$ in a fully supervised fashion, using a reconstruction loss on the target $\bar{I}$ and the prediction $\hat{I}$ with

$$\mathcal{L}_{recons} := ||\hat{I} - \bar{I}||_1. \tag{2}$$

To prevent the generator from predicting blurry outputs we adopt the perceptual loss [18]. In particular, to ensure high and low-level similarity, we extract features ϕ^l at multiple levels L from a VGG16 [43] network trained on ImageNet. We employ the first five ($|L| = 5$) different layers and calculate the perceptual loss using

$$\mathcal{L}_{perceptual} := \frac{1}{|L|} \sum_{l \in L} ||\phi^l(\hat{I}) - \phi^l(\bar{I})||_1. \tag{3}$$

We additionally use an adversarial loss to improve fine-grained reconstruction and ensure proper domain transfer. In particular, we use a discriminator which assesses if a sample in fact originates from the illumination-free domain. In our implementation we use both a global D and a local discriminator LD as proposed in [14]. While the global discriminator encourages better translation to the target domain, the local discriminator operates on small patches in order to enforce the preservation of details. We use binary cross entropy loss for both discriminators. Following common practice [14, 30, 32] we condition the output on input according to $I \oplus \hat{I}$ or $I \oplus \bar{I}$, where $\oplus$ denotes horizontal concatenation [30].

During generator training, we feed the conditioned images to both discriminators for teaching the generator to

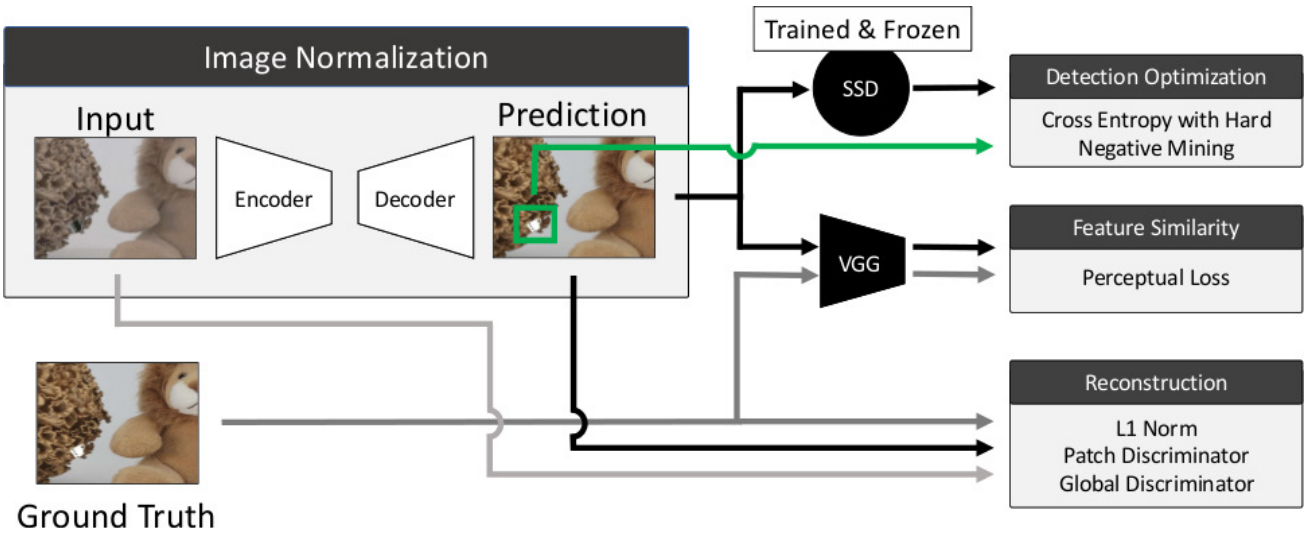

Figure 2. **Training scheme of DB-GAN for Object Detection.** Our loss is based on three different blocks, all intended to optimize detection under high lighting variations. First, a reconstruction term for high quality reconstruction of the normalized target scene $\bar{I}$. Thereby, we incorporate two discriminators ensuring consistency at different scales. Second, a perceptual term to enforce feature similarity between the prediction $\hat{I}$ and the target $\bar{I}$. Finally, a detection term in which we propagate the loss, with respect to the given ground truth (green arrow), from a pre-trained SSD instance to the image normalization network. Due to this, the network is forced to reconstruct the image $\bar{I}$ such that detection is optimized.

produce realistic images that seem to originate from the uniform lighting domain. Again, we use binary cross entropy loss for optimization. We denote these two loss term as $\mathcal{L}_{fool_D}$ and $\mathcal{L}_{fool_{LD}}$.

Unique to this work is the training of the generator with a additional detection loss (*Detection Optimization* as depicted in Fig 2). In essence, we encourage the GAN to not only create realistic illumination normalised images, but to also optimize the image for detection. To this end, we pre-train the detector on synthetic data without any illumination changes and freeze its weights. When training DB-GAN, we additionally back-propagate the loss with respect to the trained detector. DB-GAN is consequently required to adequately adjust lighting in order to optimize detection. To test out the proposed architecture, we use SSD [27] for 2D object detection and SSD6D [20] for 6D object pose estimation. For both detectors we use the original loss terms $\mathcal{L}_{Det}$, as reported in the corresponding papers. Given a set of positive Pos and hard-mined negative Neg anchor boxes, we minimize the following

$$\mathcal{L}_{Det}(\text{Pos}, \text{Neg}) := \sum_{b \in Pos} (L_{class} + \alpha L_{fit}) + \sum_{b \in Neg} L_{class}. \tag{4}$$

with respect to SSD, and for SSD6D according to

$$\mathcal{L}_{Det}(\text{Pos}, \text{Neg}) := \sum_{b \in Neg} L_{class} + \sum_{b \in Pos} (L_{class} + \alpha L_{fit} + \beta L_{view} + \gamma L_{inplane}). \tag{5}$$

Thereby L_{class} denotes the cross-entropy loss applied to each anchor and L_{fit} denotes the L1 loss which measures the misalignment of the corners in order to provide a tight fit. Further, SSD6D decouples 3D rotation into viewpoint and in-plane rotation. Thereby viewpoint describes the perceived surface and inplane rotation describes how this surface is rotated on the image-plane. To increase stability, SSD6D bins viewpoint and in-plane rotation and conducts classification referring again to the cross-entropy loss for L_{view} and $L_{inplane}$.

The final loss for the generator is then comprised of a weighted sum over all individual contributions

$$\mathcal{L} := \mathcal{L}_{recons} + \lambda_1 \mathcal{L}_{GAN} + \lambda_2 \mathcal{L}_{perceptual} + \lambda_3 \mathcal{L}_{fool_D} + \lambda_4 \mathcal{L}_{fool_{LD}} + \lambda_5 \mathcal{L}_{Det} \tag{6}$$

We empirically found that good choices for the above hyper-parameters are: $\lambda_1 = \lambda_2 = \lambda_3 = 1$, $\lambda_4 = 0.5$ and $\lambda_5 = 0.01$.

3.2. Image Enhancement Using DB-GAN

The aim of our approach is to use the trained DB-GAN to generate a new training set enhancing the detector's robustness towards different lighting conditions.

PHOS Dataset. In line with [35], we use the PHOS dataset [47] to train our DB-GAN for illumination robustness. Contrary to [35], we only use PHOS to extract background images. The PHOS dataset [47], contains 15 real

scenes, captured under 15 different lighting conditions: one *correct* exposure, 8 images under uniform lighting (*i.e*. 4 underexposed samples and 4 overexposed samples) and 6 samples with non-uniform lighting.

Baseline Detector Data Generation. As we want to back-propagate the detector loss, we need to first train a detector instance capable of detecting all objects of interest. Since we focus on training with synthetic data, we follow standard procedure [20, 29] and render 3D object models with random poses on top of random backgrounds, drawn from the Microsoft Coco dataset [26, 11]. Afterwards, we use the generated data to train the initial detector.

DB-GAN Data Generation. While the background variability in PHOS is limited, it exposes a very high per image resolution of (4256×2832). Considering the input resolution of modern deep learning architectures, this enables the sampling of numerous diverse patches. We use 256×256 as sample size, since it correlates to the input resolution of the DB-GAN generator. We use Laplacian checks to ensure only patches with sufficient textural variation are used. To generate our DB-GAN training data, we render the object models on these PHOS patches. Therefore, we randomly sample an image I from any lighting condition and utilize the matching image with the correct exposure as ground truth $\bar{I}$. We apply several light perturbations on the object model with respect to different OpenGL functionalities and render the result on I. We then re-render the same objects and poses onto the target image, however, without employing any perturbations. We demonstrate two example training examples in the supplementary material. Once the detector baseline and DB-GAN are trained, the detector training data is passed to the generator. The resulting output images form the new, normalised, training data. Finally, a new detector instance is trained on the normalised data.

3.3. Toyota TrueBlue dataset

TrueBlue is a new dataset which specifically targets to assess object detection robustness to white balance errors. Existing image datasets focusing on color temperature [5, 40, 34] do not quantify the illuminant and do not contain household objects with ground truth bounding boxes and 3D object models. We believe this dataset to be the first to be acquired with known light source color temperature and camera settings and, thus, enables quantification of detector performance under erroneous white balance conditions[1].

Toyota TrueBlue (see Figure 3) consists of 11 image sets of 3 different scenes with daily household objects with 3D model, distractor objects and also the MacBeth Color Checker chart. Each scene was illuminated from above by a set of three lights of different types, e.g. LED, incandescent, compact fluorescent, daylight and mixture of different light sources. 11 images were acquired of each scene using a Nikon D750 with Nikon 24-70mm f/2.8 lens with 11 different white balance settings, ranging from 2500K to 10000K. More details on how the dataset was acquired can be found in the supplementary material.

[1]Toyota Trueblue can be downloaded free-of-charge for non-commercial use by filling the form at https://forms.gle/ZX1aWPiu9HoetKcG9.

Figure 3. **Example of two color images from the Toyota TrueBlue dataset.** The image on the left has a 2500K color temperature, while the image on the right depicts the same scene at 10000K. More examples can be found in the supplementary material.

4. Evaluation

In this section, we introduce the implementation details, and the datasets used for evaluation. Then we demonstrate the results of our experiments.

4.1. Implementation Details

We generate 50000 training images for both the GAN and the two SSD instances and 100000 training images for SSD6D since it is a more complex task. We train all detectors for 50 epochs. Due to the different number of objects we trained DB-GAN for 10 epochs on the TUD Light objects and for 30 epochs on the Toyota Light objects, with a batch size of 1. The initial learning rate is set to 0.0003 with an exponential update rule. To stabilise training, we do not back-propagate the detector loss until the reconstructions are fairly realistic. Empirically, we found that 30000 iterations are sufficient for this. The experiments were implemented with Tensorflow [1] and run on a single Nvidia TitanXp GPU.

Generator implementation. The generator follows an encoder-decoder architecture using skip connections similar to [37]. We use a 5×5 filter size and leaky ReLU(LReLU) [56], with a 0.2 slope on the negative side, as activation function. The generator consists of eight convolutional layers with stride equal to 2 in the encoder as well as eight deconvolutional layers in the decoder. Each convolutional layer of the encoder is followed by batch normalisation. Further, the encoder has an input image size of 256×256. We use unpooling with zero padding for up-

sampling. The final up-sampling layer is followed by a hyperbolic tangent activation function to squeeze the output between 0-1 for all channels.

Discriminators implementation. The global discriminator is composed of four convolutional layers. Each convolutional layer is followed by a batch normalisation layer with LReLU activation. Finally, a fully connected layer with sigmoid activation is applied.

The local discriminator first applies a convolutional layer. Afterwards, we extract 64 non-overlapping patches of size 32×32. Each of them is processed by two more convolutional layers followed by a fully connected layer with sigmoid activation. This enforces the output to be also locally consistent within each patch.

Detectors. Our SSD and SSD6D detectors work at 299×299 resolution with an InceptionV4 backbone [45] using 6099 anchor boxes. For viewpoint classification in 6D we use 89 view vertices and 36 inplane angles.

4.2. Evaluation Protocol

To assess the performance of our method, we performed experiments on the Benchmark for 6D Object Pose Estimation (BOP) 2019 challenge version [12] of both the Toyota Light and the TUDLight datasets. For the 6D experiments on the Toyota Light dataset we trained all detectors on 4 objects, namely objects 6,9,14 and 15 that we believe well represent the dataset in terms of shape and appearance variation. Note that for all experiments we do not use any of the datasets images during training, but rather train our networks fully from synthetic renderings of the 3D model data. We compare the performance of our approach against the SSD or SSD6D baselines. We additionally compare against three illumination normalisation/image enhancement approaches: the Difference of Gaussians (DoG), EnlightenGAN[17], RetinexNet[51] and Deep Upe[48]. Among classical computer vision approaches DoG still provides top performance on image normalisation for object detection and 6D object pose estimation [35]. In our experiments we used two Gaussian kernels of size 5 and 3 pixels. For all DoG, EnlightenGAN, RetinexNet and Deep Upe we pre-processed the training dataset as well as the input images at inference time.

Finally, we show the effectiveness of our approach at increasing object detection robustness against white balance variation. To achieve this we manually perturbe the hue value of the GAN training images. The hue range $[-15, 15]$ was divided into 4 intervals of equal length. Then a random hue value was sampled in each interval and added uniformly to each image pixel producing 4 new GAN training images. The task of DB-GAN was to reconstruct the original images.

Toyota Light dataset. The Toyota Light dataset [12] contains 21 rigid household objects, captured under 5 different lighting conditions. Noteworthy, the annotation for each input sample includes the actual light conditions at the acquisition time. Two lighting levels are reported. The first is ambient light which is a diffuse overhead light source. The intensity of the incident light on the object was kept constant at 200lx for all samples. The second is the intensity of a directional light source oriented at 90 degrees to the scene. This feature makes this dataset suitable to evaluate non-uniform lighting robustness.

TUD Light Dataset. The TU Dresden Light dataset contains training and test image sequences that show three moving objects under 8 different lighting conditions. The object poses were annotated by manually aligning the 3D object model with the first frame of the sequence and propagating the initial pose through the sequence using ICP.

Metrics. All 2D experiments are evaluated following the standard metric for 2D detection, *i.e.* mean Average Precision(mAP) with a 0.5 IOU threshold. The 6D experiments are evaluated using the BOP 19 challenge toolkit. We report the average recall and report the recall according to all individual BOP metrics in the supplementary materials.

4.3. Qualitative Results

Figure 4 shows qualitative results of our approach for 2D object detection. Both in challenging dark and bright conditions DB-GAN is able to recover images that look almost identical. The SSD trained on DB-GAN generated images can detect a larger number of object instances compared to the SSD baseline. Qualitative comparisons among the different approaches are presented in the supplementary material.

4.4. Quantitative Results

Here we provide a quantitative evaluation of our detection boosting approach compared with existing works.

4.4.1 2D Object Detection

Toyota Light & TUD Light. Table 1 shows the 2D results on the Toyota Light and TUD Light datasets. Our method achieves a mAP of 0.72 on the Toyota Light dataset and 0.66 on the TUD Light dataset, outperforming the SSD baseline as well as all the other approaches. In more detail, we surpass the best existing approach by 0.43 (Deep Upe and EnlightenGAN) on the Toyota Light and by 0.04 (RetinexNet) on TUD Light.

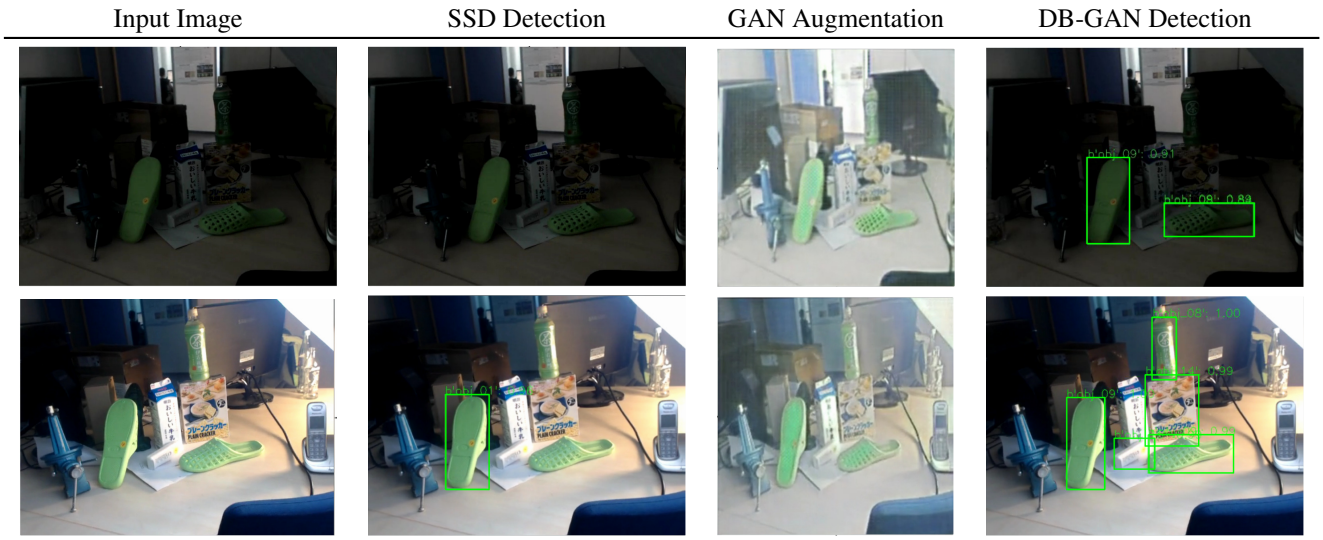

Figure 4. **Comparison of the SSD baseline with our GAN optimized SSD on objects taken from Toyota Light.** Thereby, the second column depicts the results using only SSD and the fourth column shows the corresponding detection employing DB-GAN. It can be easily deduced that our approach significantly improves detection even under difficult lightning conditions. Further, notice that almost all directional light is canceled by the GAN, as illustrated in the intermediate DB-GAN representations (3rd column).

SSD with	Toyota Light mAP ↑	TUD Light mAP ↑
DoG	0.20	0.36
enlightenGAN[17]	0.29	0.43
Retinex-Net[51]	0.28	0.62
Deep Upe [48]	0.29	0.47
baseline	0.27	0.18
DB-GAN	**0.72**	**0.66**

Table 1. **DB-GAN 2D Object Detection results on the Toyota Light and TUD Light datasets.** Our method outperforms the SSD baseline as well all other state-of-the-art approaches for illumination normalisation.

Losses used	mAP ↑
L1	0.55
+ Perceptual	0.67
+ Global Discriminator	0.66
+ Local Discriminator	0.60
+ SSD Loss	**0.72**

Table 2. **DB-GAN loss ablation study on Toyota Light.**. These results show that the best performance is achieved when using the proposed combination of loss terms.

Ablation Study. The ablation study was performed on the Toyota Light dataset for 2D object detection. We added the loss terms one by one and report the corresponding mAP. Table 2 shows the results of our ablation study with respect to each loss contribution. Noteworthy, each loss term helps to improve the overall detection performance. Importantly, our main contribution, *i.e*. the back-propagation of the detector loss, constitutes again a significant leap forward in performance, overall giving the best results.

SSD with	Toyota TrueBlue mAP ↑
baseline	0.39
baseline w/ augmentation	0.54
DB-GAN	**0.73**

Table 3. **DB-GAN results on the Toyota TrueBlue dataset.** Our method outperforms the baseline as well as SSD when leveraging color augmentations.

Toyota TrueBlue. Table 3 shows our results on color robustness. We compared our method against the SSD baseline. We additionally compared with standard color augmentation by training a SSD instance on perturbed images. In practice, we perturbed the hue channel of the training images by sampling a random value in the interval $[-15, 15]$ and adding that amount. The results show that our approach achieves a mAP of 0.73, improving on the SSD baseline by almost a factor of two and performing 0.19 better than color augmentation. The supplementary material provides visual examples of detection results for each color temperature.

4.4.2 6D Object Pose Estimation

Toyota Light & TUD Light Table 4 reports the results of our DB-GAN experiments for 6D object pose estimation.

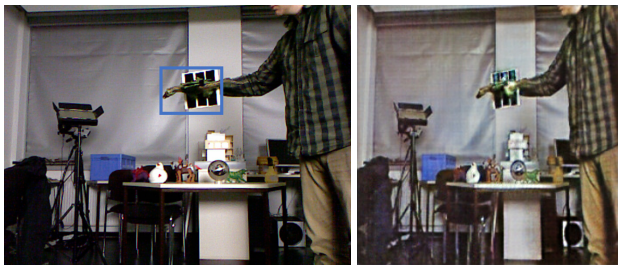

Figure 5. **Examples from the evaluation on TUD Light.** The pair shows a TUD Light image with the corresponding GAN augmentation. Notice how the GAN especially focused on the objects of interest. Nevertheless, the method is also capable of recovering structure in the background, which was almost completely lost due to bad illumination.

	Toyota Light		TUD Light	
SSD6D with	w\o ICP	w\ICP	w\o ICP	w\ICP
DoG	0.35	0.37	0.14	0.19
enlightenGAN[17]	0.30	0.34	0.157	0.21
Retinex-Net[51]	0.32	0.36	0.13	0.19
Deep Upe [48]	0.34	0.38	0.12	0.18
baseline	0.23	0.32	0.159	0.155
DB-GAN	**0.42**	**0.44**	**0.164**	**0.25**

Table 4. **Results for DB-GAN for 6D object pose estimation on Toyota Light and TUD Light.** Our method outperforms the SSD6D baseline as well all other state-of-the-art approaches for illumination normalisation.

Similar to 2D object detection, our approach improves on the baseline detector as well as all alternative approaches, by a margin of 7.2% on the Toyota Light and 0.5% on the TUD Light. Furthermore, our approach is the only one that significantly improves performance over the baseline on the TUD Light dataset. Additionally, we applied an ICP step to refine the predicted poses. Notice that our approach stays the most competitive. Furthermore, with ICP the gap between our approach and existing methods on TUD Light significantly increases.

4.4.3 Additional Experiments.

Figure 6 shows the performance of the SSD baseline and our boosted SSD on the entire Toyota Light dataset (both train and test sets) as a function of the contrast ratio. Here, contrast ratio is defined as the ratio of the intensity of incident light from the directional light source with respect to the overhead diffuse light source mentioned previously. We observe that the SSD baseline particularly struggles to detect objects in low and high contrast, while after boosting, SSD has become more light invariant, showing roughly the same level of performance for each setting. This shows that our approach is able to improve detection accuracy for both uniform lighting (Contrast Ratio=0) as well as non-uniform lighting (Contrast Ratio=1-10).

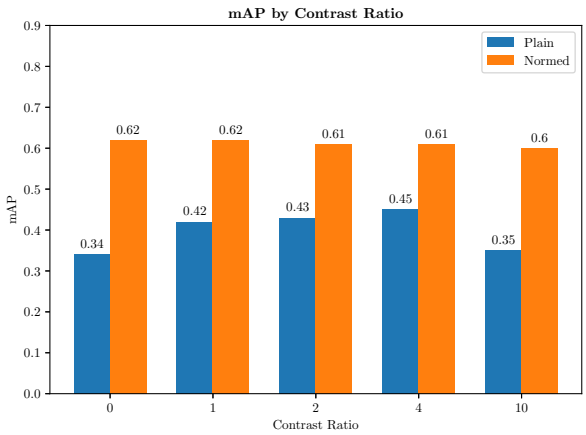

Figure 6. **Comparison between the SSD baseline and the boosted SSD with respect to different contrast ratios in the images.** We report mAP for each contrast ratio value in the Toyota Light dataset. Our approach can deal with both uniform (Contrast Ratio=0) as well as non-uniform lighting (Contrast Ratio=1-10).

DB-GAN pre-processing during inference. While pre-processing training images greatly improves performance, we also want to investigate the use of DB-GAN for pre-processing input images prior to inference. From our experiments we found that in almost all cases this further enhanced the models' capabilities. Nonetheless, when referring to Toyota Light, we surprisingly reveal a small drop in performance. Repetitive textural patterns as well as large flat areas oftentimes degrade the domain transfer capabilities of the GAN, since these samples are eminently different to our training distribution. A qualitative example is shown in Figure 5.

5. Conclusion & Future Work

We presented DB-GAN, a GAN based approach which is able to boost object detection and 6D object pose estimation performance under challenging lighting conditions. The evaluation shows that our method clearly outperforms both the baseline detectors as well as all other state-of-the-art approaches. Further, our method for image normalisation is fully data-driven and neither requires large manually annotated datasets, nor prior knowledge of the input image. Furthermore, our approach is able to deal with non-uniform lighting and does not need prior knowledge of the input image. In the future we want to explore how to expand our methods towards a more diverse set of tasks.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Yancheng Bai, Yongqiang Zhang, Mingli Ding, and Bernard Ghanem. Sod-mtgan: Small object detection via multi-task generative adversarial network. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 206–221, 2018.

[3] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[4] Yu-Sheng Chen, Yu-Ching Wang, Man-Hsin Kao, and Yung-Yu Chuang. Deep photo enhancer: Unpaired learning for image enhancement from photographs with gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6306–6314, 2018.

[5] F. Ciurea and B. Funt. A large image database for color constancy research. In *Proceedings of the Imaging Science and Technology Eleventh Color Imaging Conference*, 2003.

[6] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kehui Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.

[7] Ali Diba, Vivek Sharma, Rainer Stiefelhagen, and Luc Van Gool. Weakly supervised object discovery by generative adversarial & ranking networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[9] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference deep curve estimation for low-light image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1780–1789, 2020.

[10] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017.

[11] Stefan Hinterstoisser, Vincent Lepetit, Paul Wohlhart, and Kurt Konolige. On pre-trained image features and synthetic images for deep learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 0–0, 2018.

[12] Tomas Hodan, Frank Michel, Eric Brachmann, Wadim Kehl, Anders GlentBuch, Dirk Kraft, Bertram Drost, Joel Vidal, Stephan Ihrke, Xenophon Zabulis, et al. Bop: Benchmark for 6d object pose estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 19–34, 2018.

[13] Tomáš Hodaň, Vibhav Vineet, Ran Gal, Emanuel Shalev, Jon Hanzelka, Treb Connell, Pedro Urbina, Sudipta N Sinha, and Brian Guenter. Photorealistic image synthesis for object instance detection. In *2019 IEEE International Conference on Image Processing (ICIP)*, pages 66–70. IEEE, 2019.

[14] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.

[15] Heeoh Jang, Dongkyu Kim, Wonhyuk Ahn, and Heung-Kyu Lee. Generative object detection: Erasing the boundary via adversarial learning with mask. In *2019 IEEE 2nd International Conference on Information Communication and Signal Processing (ICICSP)*, pages 495–499. IEEE, 2019.

[16] Kevin Jarrett, Koray Kavukcuoglu, Yann LeCun, et al. What is the best multi-stage architecture for object recognition? In *2009 IEEE 12th international conference on computer vision*, pages 2146–2153. IEEE.

[17] Yifan Jiang, Xinyu Gong, Ding Liu, Yu Cheng, Chen Fang, Xiaohui Shen, Jianchao Yang, Pan Zhou, and Zhangyang Wang. Enlightengan: Deep light enhancement without paired supervision. *arXiv preprint arXiv:1906.06972*, 2019.

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.

[19] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.

[20] Wadim Kehl, Fabian Manhardt, Federico Tombari, Slobodan Ilic, and Nassir Navab. Ssd-6d: Making rgb-based 3d detection and 6d pose estimation great again. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1521–1529, 2017.

[21] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[22] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv:1811.00982*, 2018.

[23] Chengyang Li, Dan Song, Ruofeng Tong, and Min Tang. Illumination-aware faster r-cnn for robust multispectral pedestrian detection. *Pattern Recognition*, 85:161–171, 2019.

[24] Zhigang Li, Gu Wang, and Xiangyang Ji. Cdpn: Coordinates-based disentangled pose network for real-time

rgb-based 6-dof object pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7678–7687, 2019.

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[26] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[27] Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.

[28] Fabian Manhardt, Diego Martin Arroyo, Christian Rupprecht, Benjamin Busam, Tolga Birdal, Nassir Navab, and Federico Tombari. Explaining the ambiguity of object detection and 6d pose from visual data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 6841–6850, 2019.

[29] Fabian Manhardt, Wadim Kehl, Nassir Navab, and Federico Tombari. Deep model-based 6d pose refinement in rgb. In *The European Conference on Computer Vision (ECCV)*, September 2018.

[30] Mehdi Mirza and Simon Osindero. Conditional generative adversarial nets. *arXiv preprint arXiv:1411.1784*, 2014.

[31] Sean Moran, Pierre Marza, Steven McDonagh, Sarah Parisot, and Gregory Slabaugh. Deeplpf: Deep local parametric filters for image enhancement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12826–12835, 2020.

[32] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2642–2651, 2017.

[33] Kiru Park, Timothy Patten, and Markus Vincze. Pix2pose: Pixel-wise coordinate regression of objects for 6d pose estimation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7668–7677, 2019.

[34] Andrew Blake Tom Minka Toby Sharp Peter Gehler, Carsten Rother. Bayesian color constancy revisited. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2008.

[35] Mahdi Rad, Peter M. Roth, and Vincent Lepetit. Alcn: Adaptive local contrast normalization for robust object detection and 3d pose estimation. In *BMVC 2017*, 2017.

[36] Joseph Redmon and Ali Farhadi. Yolov3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018.

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

[38] Dimitrios Sakkos, Edmond SL Ho, and Hubert PH Shum. Illumination-aware multi-task gans for foreground segmentation. *IEEE Access*, 7:10976–10986, 2019.

[39] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4570–4580, 2019.

[40] Lilong Shi and Brian Funt. Re-processed version of the gehler color constancy dataset of 568 images, 2010.

[41] Xuepeng Shi, Shiguang Shan, Meina Kan, Shuzhe Wu, and Xilin Chen. Real-time rotation-invariant face detection with progressive calibration networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2295–2303, 2018.

[42] Oleksii Sidorov. Conditional gans for multi-illuminant color constancy: Revolution or yet another approach? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019.

[43] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[44] Martin Sundermeyer, Zoltan-Csaba Marton, Maximilian Durner, Manuel Brucker, and Rudolph Triebel. Implicit 3d orientation learning for 6d object detection from rgb images. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 699–715, 2018.

[45] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[46] Jonathan Tremblay, Thang To, and Stan Birchfield. Falling things: A synthetic dataset for 3d object detection and pose estimation. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2038–2041, 2018.

[47] Vasillios Vonikakis, Dimitrios Chrysostomou, Rigas Kouskouridas, and Antonios Gasteratos. A biologically inspired scale-space for illumination invariant feature detection. *Measurement Science and Technology*, 24(7):074024, 2013.

[48] Ruixing Wang, Qing Zhang, Chi-Wing Fu, Xiaoyong Shen, Wei-Shi Zheng, and Jiaya Jia. Underexposed photo enhancement using deep illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6849–6857, 2019.

[49] Wanwei Wang, Wei Hong, Feng Wang, and Jinke Yu. Gan-knowledge distillation for one-stage object detection. *IEEE Access*, 8:60719–60727, 2020.

[50] Xiaolong Wang, Abhinav Shrivastava, and Abhinav Gupta. A-fast-rcnn: Hard positive generation via adversary for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2606–2615, 2017.

[51] Chen Wei, Wenjing Wang, Wenhan Yang, and Jiaying Liu. Deep retinex decomposition for low-light enhancement. *arXiv preprint arXiv:1808.04560*, 2018.

[52] Yu Xiang, Tanner Schmidt, Venkatraman Narayanan, and Dieter Fox. Posecnn: A convolutional neural network for 6d object pose estimation in cluttered scenes. *arXiv preprint arXiv:1711.00199*, 2017.

[53] Kai-Fu Yang, Xian-Shi Zhang, and Yong-Jie Li. A biological vision inspired framework for image enhancement in poor

visibility conditions. *IEEE Transactions on Image Processing*, 29:1493–1506, 2019.

[54] Qing Zhang, Yongwei Nie, Lei Zhu, Chunxia Xiao, and Wei-Shi Zheng. Enhancing underexposed photos using perceptually bidirectional similarity. *IEEE Transactions on Multimedia*, 2020.

[55] Shanshan Zhang, Jian Yang, and Bernt Schiele. Occluded pedestrian detection through guided attention in cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6995–7003, 2018.

[56] Xiaohu Zhang, Yuexian Zou, and Wei Shi. Dilated convolution neural network with leakyrelu for environmental sound classification. In *2017 22nd International Conference on Digital Signal Processing (DSP)*, pages 1–5. IEEE, 2017.

[57] Yang Zhang, Changhui Hu, and Xiaobo Lu. Il-gan: Illumination-invariant representation learning for single sample face recognition. *Journal of Visual Communication and Image Representation*, 59:501–513, 2019.

[58] Kun Zheng, Mengfei Wei, Guangmin Sun, Bilal Anas, and Yu Li. Using vehicle synthesis generative adversarial networks to improve vehicle detection in remote sensing images. *ISPRS International Journal of Geo-Information*, 8(9):390, 2019.

[59] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.