Expert Systems With Applications 224 (2023) 120045

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

Shadows & Lumination: Two-illuminant multiple cameras color constancy dataset

Ilija Domislović *, Donik Vršnak, Marko Subašić, Sven Lončarić

Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia

ARTICLE INFO

Dataset link: http://bit.ly/shal_dataset

Keywords:
Multi-illuminant dataset
Color constancy
Image color analysis
Image processing
Image segmentation

ABSTRACT

In this paper, we introduce a new large-scale publicly available color constancy dataset which we are calling the Shadows & Lumination dataset. The dataset contains 2500 minimally processed images from various indoor, outdoor, and night-time scenes. This dataset is GDPR-compliant, as we masked out all sensitive private information from the images. Unlike most other color constancy datasets, our dataset contains real-world images with two illuminants is appropriate for multi-illuminant estimation. In addition to the illumination, we provide a binary segmentation mask for each image. In the segmentation mask, we divide the image into two regions, where each region is illuminated by only one of the illuminants. We give an explanation of the methodology used to create the dataset. For dataset creation, we used five cameras: Canon 5D, Canon 550D, Sony α300, Panasonic FZ1000, and the Motorola one fusion+ mobile camera. Finally, we tested several state-of-the-art illumination estimation and image segmentation models on our dataset. The dataset is publicly available[1]. This paper also benchmarks several illumination estimation methods as well as several image segmentation methods on our dataset.

1. Introduction

Light source chromaticity has a noticeable influence on the color of an object illuminated by a light source. A crucial feature of an object is its color and the human visual system (HVS) developed the ability to perceive the intrinsic color of an object even when that scene illumination alters its color (Fairchild, 2013). The intrinsic color of an object is present when the object is illuminated by the canonical illuminant, which is usually a perfectly white light.

The fact that an object's color changes depending on what light source illuminates it presents a significant problem for many computer vision tasks (Wang et al., 2019). The object's color is an essential feature, and computer vision tasks such as object detection and object tracking assume that a red apple will be red no matter the illumination. The process of removing the illumination chromaticity in an image is also called computational color constancy. Many different methods exist to solve this problem, from simple approaches (Van De Weijer et al., 2007) that use image statistics to complex (Hu et al., 2017) ones that employ neural networks.

Methods based on convolutional neural networks (CNN) achieve the best results (Xiao et al., 2020). Neural networks need a diverse set of samples to produce accurate results. Datasets (Cheng et al., 2014; Laakom et al., 2021) with many images do exist, but these datasets contain images with only one uniform illuminant. Consequentially, methods trained on these datasets assume that a scene is illuminated by one uniform illuminant, which is not always accurate. An example of an image with two illuminants is an outdoor daytime image where a part of the scene is in the shadows. In such a situation, one illuminant is the sun, and the other is the sky in the regions which the sun does not reach. There are datasets with images that contain more than one illuminant (Beigpour et al., 2013; Bleier et al., 2011; Gijsenij, Lu et al., 2011), but these datasets usually lack variety and contain only a small number of images.

In this paper, we introduce a large-scale multi-illuminant dataset that contains 2500 images with two illuminants. For each image, we provide a file containing the extracted illuminants and a segmentation mask file. The segmentation mask divides the image into two regions, and each region is illuminated by only one illuminant. The dataset contains images of real-world scenes. In the dataset, there are a variety of daytime, night-time, indoor, and outdoor scenes. Images in the dataset are processed to be GDPR-compliant, as all sensitive private

* Corresponding author.
E-mail addresses: ilija.domislovic@fer.hr (I. Domislović), donik.vrsnak@fer.hr (D. Vršnak), marko.subasic@fer.hr (M. Subašić), sven.loncaric@fer.hr (S. Lončarić).

[1] bit.ly/shal_dataset Direct link

https://doi.org/10.1016/j.eswa.2023.120045
Received 14 June 2022; Received in revised form 14 February 2023; Accepted 1 April 2023
Available online 5 April 2023

information in each image is masked out. The dataset was created using five different cameras, four professional cameras, and a mobile phone camera. This dataset is a continuation of the research done in a previously published paper (Domislović et al., 2021).

This paper consists of the following sections. Section 2 provides a formal definition of the problem. Section 3 is an overview of the currently available single and multi-illuminant datasets. Section 4 gives an overview of the dataset and insight into how images were collected and labeled, as well as the structure of the dataset. Section 5 aims to explain how the dataset should be used for the purposes of method evaluation. Section 6 presents the results obtained using existing illumination estimation and image segmentation methods. Section 7 provides the conclusion of the paper.

2. Color constancy overview

The Human Visual System is very adaptable, as we can perceive an object's intrinsic color even when the color is altered by the scene illumination, whereas computers need to perform white-balancing to remove the effect of the illuminant on an object's color. The process of white-balancing can be divided into two steps, the first step being the estimation of the image illumination image scene, and the second step being the chromatic adaptation of the image using the estimated illumination.

2.1. Problem formulation

An image is made out of pixels, where each pixel has three channels that represent the red, green, and blue intensity of the pixel $\mathbf{f} = (f_r, f_g, f_b)$. An image formation can be represented using the Lambertian model (Gijsenij, Gevers et al., 2011).

$$f_c = \int_{\omega} I(\lambda) S(\mathbf{x}, \lambda) p_c(\lambda)\, d\lambda \tag{1}$$

Pixel RGB intensities depend on illumination color $I(\lambda)$, surface reflectance $S(\mathbf{x}, \lambda)$ and the camera sensitivity function $\mathbf{p}(\lambda) = (p_r(\lambda), p_g(\lambda), p_b(\lambda))$ of the three channels, where $c = \{r, b, g\}$, ω represents the visible spectre, $\mathbf{x}$ the spatial coordinates, and λ the light wavelength.

The second step, chromatic adaptation, is usually performed using the von Kries model (von Kries, 1905). This model assumes that the red, green, and blue sensor responses are independent, and is represented using a diagonal matrix. It was shown that a diagonal matrix is sufficient for chromatic adaptation (Finlayson et al., 1993).

$$I^c = \Lambda^{u,c} * I^u \tag{2}$$

I^c is the image taken under the canonical illuminant, $\Lambda^{u,c}$ represents the von Kries diagonal matrix, and I^u is the image taken under an unknown illuminant. The diagonal matrix can be expressed as:

$$\Lambda^{u,c} = \begin{bmatrix} L_r^c/L_r^u & 0 & 0 \\ 0 & L_g^c/L_g^u & 0 \\ 0 & 0 & L_b^c/L_b^u \end{bmatrix} \tag{3}$$

where L_r^u, L_g^u, L_b^u are the red, green, and blue values of the unknown illuminant and L_r^c, L_g^c, L_b^c are the red, green, and blue values of the canonical illuminant. The canonical illuminant is the white light with a $\mathbf{L}^c$ value of $(1, 1, 1)^T$.

This process is used when an image has only one illuminant and is not easy to execute when multiple illuminants are present in the image. The illumination of each pixel is needed to perform the von Kries (1905) chromatic image adaptation. For this reason, the images in our dataset have been taken so that the illumination is uniform in each illuminant region, with a clearly defined border between illumination regions.

3. Previously published color constancy datasets

One of the oldest still widely used single illuminant datasets is the ColorChecker (Gehler et al., 2008) dataset introduced in 2008. This dataset contains over 500 images created using two cameras and the illumination was extracted using the Macbeth ColorChecker which has gray, white, and chromatic surfaces used to extract scene illumination. The dataset has both indoor and outdoor images. Improper illumination extraction and confusing wording on image black-level subtraction in Gehler et al. (2008) resulted in three different sets of illumination ground truths (Finlayson et al., 2017). The existence of three ground truths also causes problems when comparing methods since it was shown in Finlayson et al. (2017) that the used ground truth can significantly affect method accuracy.

Another older single illuminant dataset is the NUS-8 (Cheng et al., 2014) dataset. It contains over 1800 images taken by 9 cameras. The ColorChecker was also used to extract the illumination of the images in the dataset. Around 210 images were taken using each camera. The dataset contains both indoor and outdoor images and involves the largest number of cameras out of any existing dataset. However, the problem with this dataset is that it does not contain over 1800 unique images. Instead, it contains only around 210 different scenes. Each scene was captured by multiple cameras. This dataset is often used to see how well the methods perform on unknown cameras.

The largest single illuminant dataset is the Intel-TAU (Laakom et al., 2021) dataset, which contains over 7000 images. The dataset was created with three different cameras and contains indoor scenes, outdoor scenes, laboratory printout scenes, and laboratory environment scenes. The dataset also uses the ColorChecker to extract illumination ground truth.

Another dataset is the Cube+ (Banić et al., 2017) dataset. This dataset contains over 1700 single illuminant images all of which have been captured by a single camera, containing indoor, daytime outdoor, and night-time outdoor scenes. Unlike the previously mentioned datasets, the authors used a SpyderCube to extract the illumination and unlike the ColorChecker, the SpyderCube only has gray and white surfaces that can be used for illumination extraction.

As the main focus of computational color constancy was illumination estimation in images with single uniform illumination, there are not as many multi-illuminant datasets. Existing multi-illuminant datasets (Beigpour et al., 2013; Bleier et al., 2011; Gijsenij, Lu et al., 2011) are fairly small, containing less than 100 images.

An example of a multi-illuminant dataset is described in Bleier et al. (2011). It is a fairly small dataset containing 36 images that were captured using a Canon EOS 550D camera. Four different scenes were captured in 9 different multi-illuminant environments. Both the scene and illumination conditions were artificially created in a laboratory environment. The illumination setup was created using two Reuter lamps with LEE color filters.

The second dataset is the Multiple Light Sources (Gijsenij, Lu et al., 2011) dataset. This dataset holds almost twice as many images as the dataset from Bleier et al. (2011). Images were taken using the Sigma SD10 camera. In addition to laboratory environment images(59), this dataset also contains real-world images(9). In order to extract the illumination, gray objects were placed in the scenes.

The last on our list is the Multiple-Illuminant Multi-Object (Beigpour et al., 2013) dataset. It is the largest of the presented multi-illuminant datasets, containing 80 images. The images were captured using the Sigma SD10 camera as well. The dataset consists of 60 images taken in a laboratory environment and 20 real-world images, but does not contain 80 unique images. Instead, the 60 laboratory images are 10 different scenes that were taken under 6 different illumination conditions.

In addition to the dataset, the authors of Beigpour et al. (2013) also propose a method for the automatic creation of a per-pixel illumination mask. This mask shows the influence of each illuminant on each pixel. To achieve this, the authors took three images of each scene: one with

Table 1
Table representing the number of images taken by each camera presented by type of image.

	Outdoor	Indoor	Nighttime
Canon 5D	395	39	61
Canon 550D	403	44	57
Motorola	400	40	59
Sony	400	38	60
Panasonic	395	39	70

both illuminants, one with only the first illuminant, and one with only the second illuminant present. To create the mask, they used these images as well as the fact that the scene taken under both illuminants is the sum of the two scenes where only one of the illuminants is present. This method was not used in this paper, as for this method to work, we would need to be able to turn off scene illumination. This is a very restrictive condition since the illumination cannot be turned off or entirely obscured in most real-world situations, for which the most obvious example is the sun.

4. Dataset overview

In this paper, we introduce the Shadows & Lumination dataset, a large-scale publicly available multi-illuminant dataset. The images in the dataset are minimally processed. Images are processed with `dcraw` an open-source program that decodes various RAW image formats into standard, commonly used image formats. The flags `-T -D -4` are used. The program outputs an image in a 16-bit format in the camera's RAW color space.

Afterward, simple debayering was performed. The red and blue components were directly taken from the Bayer pattern and the green component was obtained by averaging the two green components of the Bayer pattern. This method reduces the width and height of the image by half. Since RAW images have very high dimensionality, this does not have a noticeable effect on how an image looks. We used this method as it has almost no artifacts to create minimally processed images for our dataset. Some images also contain black boxes over parts of the image. These black boxes are used to mask sensitive private information, such as faces, to make the dataset GDPR-compliant.

The camera setup used to take images differs from image to image. There are no consistencies between the images except for ISO, where the smallest ISO was used. For the sake of completeness, we provide the EXIF metadata with each image.

The dataset contains 2500 images from a diverse set of scenes taken in various locations. The images in the dataset can be divided into groups based on what scene the image captures. There are three types: images taken outside during the daytime, images taken outside during the night-time, and images that were taken indoors during various times of the day. Five different cameras by four manufacturers were used to create the dataset. Cameras used were the Canon EOS 5D Mark II, the Canon EOS 550D, the Panasonic DMC-FZ1000, the Sony DSLR-α300, and the Motorola One Fusion+ mobile phone camera. Motorola One Fusion+ uses the Samsung ISOCELL Plus GW1 1/1.72" camera sensor. All scene types were captured with each camera to keep the dataset balanced. The exact numbers can be seen in Table 1. Example images from all cameras can be seen in Fig. 1.

4.1. Dataset creation

Each image in the dataset is accompanied by the extracted ground truth illumination and a segmentation mask that divides the image into two regions. Each region contains only one illuminant.

4.1.1. Illumination

The illuminants in the dataset can be divided into two groups. Each image has one illuminant from each group. The first is a direct light source and the second is an ambient light source. A simple example of a direct light source is the sun, whereas a clear sky is a simple example of an ambient light source. The direct light source is the stronger illuminant that only illuminates part of the scene. The ambient light source illuminates the entire image, but only has an effect in the regions where the direct light does not reach since it is the weaker illuminant and gets suppressed by the stronger illuminant.

As mentioned previously, in the daytime image, the two illuminants are the sun and the sky. For night-time and indoor images, the situation is slightly more complex. In night-time images, the direct light source is most often a LED, incandescent, or sodium street lamp. The night sky can be seen as the ambient light, but the problem is that it has a low intensity, and other artificial lights often overpower it. This led to instances where there is a clear shadow in the image, but its illumination color is almost identical to the direct light source. The ambient light is not well-defined and is often a combination of artificial lights in the area reflected from various surfaces. This also causes the creation of images where the shadow contains non-uniform illumination that is difficult to segment. Such images were discarded as they do not satisfy the requirements of the dataset. A tungsten lightbulb is the direct light source in most indoor images. Here the ambient light is also not well-defined and the same problems as in night-time images can occur. How these images were detected and removed is explained in more detail later in this section.

To extract the illumination, the SpyderCube calibration object was used. A SpyderCube has four different faces. Two faces are white(WL, WR) and two are spectrally neutral 18% gray faces(GL, GR). The faces on the SpyderCube are divided into two surfaces, left (GL, WL) and right (GR, WR) which are positioned at an angle. Each surface contains one gray and one white face. A SpyderCube can be seen in Fig. 2. The illumination was extracted from gray faces by calculating the average value of a gray face. White faces were not used as illumination, since they were oversaturated in some images.

The extracted illuminations can be seen in Figs. 3 and 4. Fig. 3 shows how the illuminants are distributed based on which camera was used to capture the image. Fig. 3 shows that the two Canon cameras have similar illumination distribution, which makes sense since they are from the same manufacturer. But Fig. 3 also shows that the Sony and Motorola cameras have similar illumination distributions, even though the Sony camera uses a Sony camera sensor and the Motorola uses a Samsung camera sensor.

Fig. 4 displays how the illuminants are distributed based on image type. It shows that the outdoor and night-time images have significantly different illuminations, while indoor image illuminations are in-between outdoor and night-time illuminations since indoor images contain both natural lighting and artificial lighting.

Since SpyderCube is a simple and relatively cheap calibration device, we performed experiments to test SpyderCube's precision, primarily to see how consistent the extracted illuminant is between different SpyderCubes. The test was executed using three SpyderCubes to detect the similarity between the extracted illuminants. In the experiment, we placed the SpyderCubes in such a way that all of them were illuminated by the same illuminant. The illuminations were extracted from the gray faces of each SpyderCube and were subsequently grouped into left and right based on which side of the SpyderCube they were, respectively. The similarities between all the faces in a group were calculated using angular distance. The side with the smaller angular distance between the faces was selected. The angular distance is calculated using Eq. (4), where $\mathbf{L}$ and $\hat{\mathbf{L}}$ are illumination vectors extracted from two SpyderCube faces. The illumination vector is obtained by calculating the average RGB value of the SpyderCube face.

After selecting a side, the angular distance between the gray-face illuminants was compared with the illuminants extracted from the

Fig. 1. A couple example images. The leftmost image was taken by the Motorola camera. The top row images were taken by the Canon 5D and Canon 550D cameras. The bottom row images were taken by the Sony and Panasonic cameras.

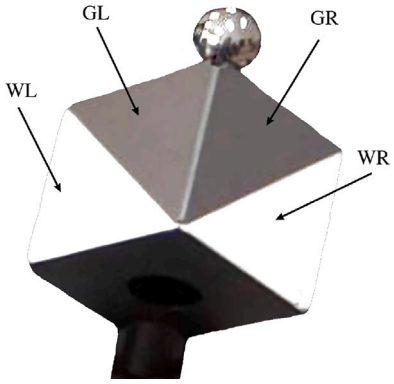

Fig. 2. A SpyderCube. GL and GR represent the gray faces. WL and WR represent the white faces.

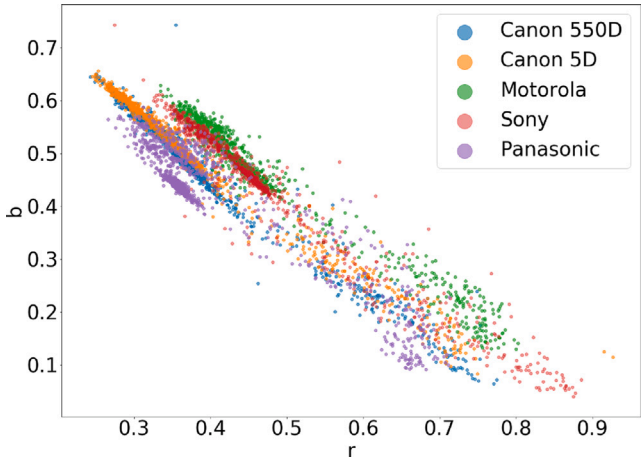

Fig. 3. Illumination distribution of the dataset by camera.

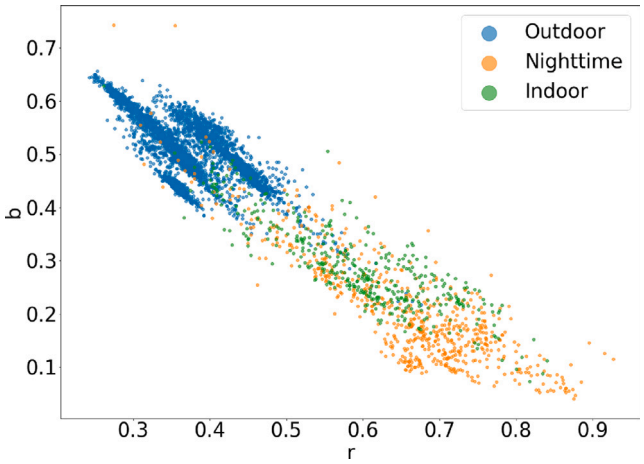

Fig. 4. Illumination distribution of the dataset by image type.

white faces on the same surface. The gray face with the smallest angular distance to its white face was selected as the ground truth illumination. This experiment was performed for direct and ambient light sources.

The test was done on 170 images. It showed that in 20% of images, the angular distance of the extracted illuminants was over 2° for the ambient illuminant. The average angular distance of the images is 3.05°. The experiment revealed that the variety of reflective surfaces, the orientation of the SpyderCubes, and the distance of the SpyderCubes from the camera, as well as to one another could all affect the ambient illumination extracted from the SpyderCubes. Because of this, the ambient illumination needs an extra SpyderCube to confirm that the extracted illumination is precise. The angular distance of 2° was selected as the border between one and two illuminants based on research done in Hordley (2006).

For direct illumination, the difference between cubes was not greater than 1°. Due to the aforementioned experiments, each image contains three SpyderCubes, one for the direct light source and two for the ambient light source. The angular distance between the SpyderCubes in ambient light is less or equal to 2°, while the angular distance between the SpyderCubes in direct light and ambient light is always greater than 2°.

These experiments demonstrate that even though the regions in the shadows appear as though they are uniformly illuminated, there are situations where the shadow illumination map can be non-uniform and may contain multiple illuminants. Consequently, a more rigorous method for illumination extraction is required so that we can be sure that the illumination in the shadows is uniform.

The face selection for the direct illuminant was simple since only one SpyderCube is used. In most situations, only one gray face was illuminated by the direct light source, and that face was selected for illumination extraction. In situations where both faces were illuminated by the direct light source, the gray face which is more similar to its white face by the measure of angular distance was selected as the ground truth. In the final situation, where both gray faces are illuminated by the direct light source and the white faces are oversaturated,

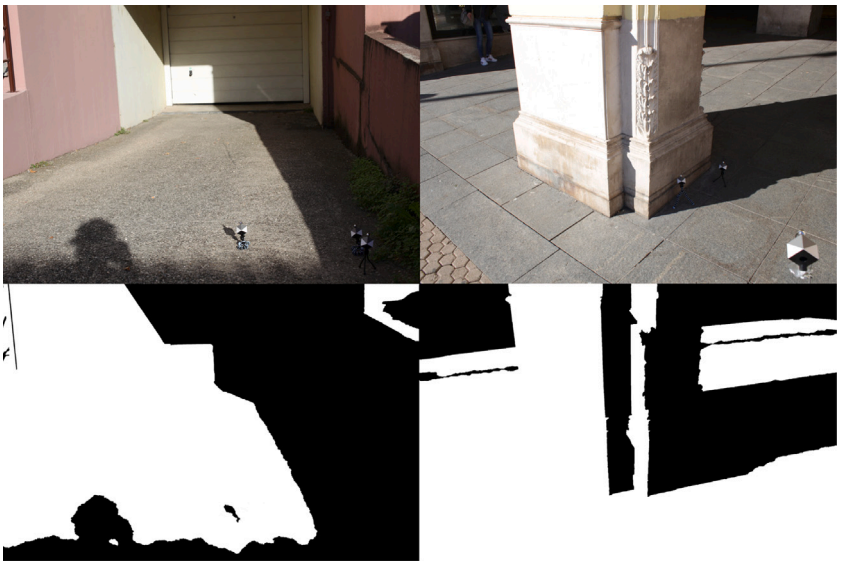

Fig. 5. Two images and their segmentation masks underneath them.

the angular distance between the gray faces is calculated. If the distance was less than 2°, the left face was used as the ground truth. If the distance was greater than 2°, the image was removed from the dataset.

The face selection for ambient light was similar. The illumination was extracted from the gray faces on both SpyderCubes and the gray faces were then grouped into left and right. The angular distance between the gray faces in each group was calculated and the group with the smaller angular distance was then selected. In the selected group, the angular distance between each gray face and its white face was calculated. The gray face with the smaller angular distance from its white face was finally selected as the ground truth.

The use of two SpyderCubes in ambient light and one in direct light does not help us during image acquisition, but it does allow us to filter out images with improper illumination labels from the dataset. It allows us to ensure that the two illuminants present in the image are different enough for the image to have multi-illuminant lighting. We used 2° as the threshold. The research from Hordley (2006) states that humans can differentiate two colors if their angular distance is over 2°. Finally, it allows us to confirm that illumination in the shadows is uniform so that the image can be used for proper multi-illuminant estimation.

4.1.2. Segmentation mask

The other part of the dataset is the illumination segmentation masks. An illumination segmentation mask is notoriously difficult to create when an image has multiple illuminants. A method for automatic labeling has been proposed (Beigpour et al., 2013), but the problem with this method is that we need to be able to turn off present illuminants. This is not possible in most situations. Therefore, we simplified the labeling process. Because each image has a direct light source and an ambient light source, the illuminants are uniform and there is a clear border between them. This makes the labeling processing simpler, but the labeling must still be done manually. This very time-consuming process was performed by multiple experts, resulting in 2500 illumination segmentation masks.

Since the labeling is done manually, the chances of subjective errors are much higher. To minimize this error, each image was corrected using the extracted illumination and the created segmentation mask. Multiple experts were then used to validate that the correct illuminant was used and that the segmentation mask regions coincide with the illumination regions of each illuminant. A couple of example images and their masks can be seen in Fig. 5.

4.2. Dataset structure

The dataset contains five folders, one for each camera. Each camera folder contains three folders: one for outdoor daytime, one for outdoor night-time, and one for indoor images. Each image has its own folder which is indexed starting at 1. For each image, we provide the minimally processed `png` file, the ground truth illumination file, the `png` illumination segmentation mask file, the `txt` file containing calibration object locations, the `txt` file containing the image regions where the sensitive data has been masked, and the image EXIF metadata text file. The ground truth illumination file contains the L2 normalized RGB values of the two illuminants, the first value being the ambient illuminant and the second being direct light. The segmentation mask file contains a binary mask where 1 represents pixels under the ambient illuminant and 2 represents pixels under the direct illuminant. The calibration object file contains the locations of the calibrations objects used to extract the scene illumination. This file is present, as the calibration object needs to be removed before using the image for neural network training since a neural network can be trained to search for the calibration object and extract the ground truth from it. The EXIF metadata is provided instead of RAW images since RAW images are not GDPR-compliant.

A couple of preprocessing steps need to be applied to properly use the dataset. Firstly, the blacklevel needs to be removed from the image. Different cameras have different blacklevels. The Sony camera has no blacklevel. The Motorola camera has a blacklevel of 63. The Panasonic camera has a blacklevel of 127. The Canon 5D has a blacklevel of 1024. The Canon 550D has a blacklevel of 2048. Secondly, the calibration objects must be masked. Finally, the oversaturated pixels need to be set to 0. We provide a Python script image loader that performs the preprocessing steps with the dataset.

5. Dataset evaluation

For proper dataset evaluation, we propose three different evaluation protocols. The first protocol uses all the images to evaluate a method. The second evaluates how well a method performs when it encounters images from an unknown camera sensor. The final protocol tests how well a method performs on different types of images.

5.1. Metrics

We used the angular error metric to evaluate the results of multi-illuminant estimation methods.

$$Angular\ error = cos^{-1}\left(\frac{\mathbf{L}\cdot\hat{\mathbf{L}}}{\|\mathbf{L}\|_2\|\hat{\mathbf{L}}\|_2}\right) \tag{4}$$

$\mathbf{L}$ represents the ground truth RGB illumination vector, $\hat{\mathbf{L}}$ represents the predicted RGB illumination vector, $\cdot$ the scalar product, and $\|\ \|$ the L2 norm. The angle is calculated in degrees.

The dice metric was used to evaluate the results of the multi-illuminant image segmentation methods.

$$Dice = \frac{2TP}{2TP + FP + FN} \tag{5}$$

TP are true positives or pixels that have correctly been labeled as ambient illuminant pixels, FP are false positives or pixels that have incorrectly been labeled as ambient illuminant pixels, and FN are false negatives or pixels that have incorrectly been labeled as direct illuminant pixels.

5.2. Use-All protocol

The first protocol is the simplest one and is commonly used for other color constancy datasets. For this protocol, all the images in the dataset are used. We propose a 5-fold split. Each fold contains around 500 images. All cameras and all image types are present in each fold. We provide the results obtained by taking the average of the results from the five folds.

5.3. One-to-Many protocol

The next protocol tests how well a method performs when it encounters a camera sensor not seen during training. For this protocol, the dataset is again divided into five subsets. This time, each fold contains images from only one camera. Each camera is used for training once, while being used for testing four times. We provide the average result of all five experiments.

5.4. By-Type protocol

The final protocol represents three different experiments. With this protocol, we test how well a method performs on the three types of images present in the dataset. The three types are daytime outdoor, night-time outdoor, and indoor images. For each of the three subsets, we propose a 5-fold split. The splits are provided with the datasets. Like the Use-All protocol, one 5-fold subset is used for testing while the others are used for training. We provide the results for each of the image type subsets.

6. Experimental results

In addition to the dataset, we also provide results obtained using existing methods. Two types of experiments were performed on the dataset. The first is image segmentation, in order to determine how existing image segmentation models perform when the task is to segment the image into regions where only one illuminant is present. The second is illumination estimation to establish how existing multi-illuminant estimation models perform on this dataset. For the sake of completeness, we include traditional non-learning methods and more complex learning-based methods for dataset evaluation.

To validate each method we use a 5-fold split, which means there are 5 model instances and 5 testing sets. The errors from all the test sets were then grouped and used to calculate the metrics which are shown in the tables. We used this approach because we get more accurate results for the best 25%, worst 25%, median, and trimean than calculating the metrics for each fold and then averaging the results.

6.1. Image segmentation results

For the traditional non-learning-based methods, we used the most popular methods from the literature. These methods are simple thresholding, Max Entropy (Leung & Lam, 1994), and OTSU (Otsu, 1979).

For simple thresholding, we transform the image into HLS color space and create a mask by selecting a threshold level and making a binary mask from the L channel. For OTSU, we transform the image into grayscale color space and select a threshold so that the variance for each of the created classes is minimized. Finally, for Max Entropy the image is also transformed into grayscale color space, but the selected threshold maximizes the entropy of the data in each class.

For learning-based methods, we experimented with four widely used image segmentation models. These models are used in a variety of different computer vision tasks, and we decided to test them on the problem of image segmentation based on illumination. As the backbone of these models, we decided to use two popular models: VGG16 (Simonyan & Zisserman, 2014) and SEResNet18 (Hu et al., 2018). We employ transfer learning (Tan et al., 2018) with backbones pretrained on ImageNet (Deng et al., 2009). The backbones are used for image feature extraction. We use these backbones in combination with the UNet, LinkNet, PSPNet, and FPNNet segmentation models. The main feature of these segmentation models is that they are made of two parts, the encoder, and the decoder. The encoder extracts useful image information by downsampling using convolutional layers, and the decoder upsamples the extracted information using deconvolutional layers. They use skip connection to combine information extracted at each downsampling stage with its corresponding upsampling stage. In UNet the stages are combined by concatenating the downsampling and upsampling stages. In LinkNet the stages are combined by adding the downsampling and upsampling stages. FPNNet also combines the upsampling and downsampling stages by addition, but they have the same number of filters for each deconvolutional layer in the decoder. They also perform predictions independently for each upsampling stage. PSPNet uses a pyramid pooling module which is placed between the encoder and decoder. This module uses convolutional layers of different kernel sizes to extract global and local image information.

The results of the multi-illuminant image segmentation can be viewed in Table 2. The mean dice score for all protocols can be observed here. Table 2 shows that the learning-based methods outperform the traditional methods. The worst performing method is Max-Entropy (Leung & Lam, 1994) which has the worst results for all the protocols. The results of the traditional methods on the Use-All and One-to-Many are the same since these methods require no training, so the two protocols do not affect them.

For the Use-all protocols, the learning-based methods have very similar performance with most models having a dice score of around 0.88, with FPN (Lin et al., 2017) having the best dice score of 0.893. The results on the One-to-Many protocol are slightly worse than the Use-all protocol, with dice scores of around 0.85. The unknown camera sensors in the testing set and the smaller training set cause the lower model accuracy. For this protocol, the best performance is obtained by UNet (Ronneberger et al., 2015).

When looking at the By-Type protocol, we can see that the easiest images to segment are the Outdoor images and the hardest to segment are the Indoor images. This makes sense since Outdoor images contain only natural light, while Indoor images contain both artificial and natural lighting. The Table also shows that images with only artificial lighting are harder to segment than images with only natural lighting. The dice score for Outdoor images is around 0.89, the dice score for Nighttime images is around 0.86, and the dice score for Indoor images is around 0.79. UNet performs best on Outdoor images, while FPN performs the best on Indoor and Nighttime images.

Table 2
The mean and standard deviation dice scores of different methods tested using different protocols. The results with the best mean are bolded.

Method	Use-All	One-to-Many	Outdoor	Indoor	Nighttime
Simple thresholding	0.809 ± 0.126	0.809 ± 0.126	0.808 ± 0.128	0.736 ± 0.149	0.802 ± 0.153
Max entropy (Leung & Lam, 1994)	0.715 ± 0.224	0.715 ± 0.224	0.730 ± 0.202	0.656 ± 0.242	0.684 ± 0.289
Otsu (1979)	0.801 ± 0.107	0.801 ± 0.107	0.809 ± 0.106	0.748 ± 0.096	0.779 ± 0.118
UNet (VGG16) (Ronneberger et al., 2015)	0.882 ± 0.120	0.851 ± 0.133	0.888 ± 0.120	0.791 ± 0.130	0.864 ± 0.128
UNet (SEResNet18) (Ronneberger et al., 2015)	0.892 ± 0.125	**0.868 ± 0.128**	**0.901 ± 0.122**	0.801 ± 0.114	0.874 ± 0.123
LinkNet (VGG16) (Chaurasia & Culurciello, 2017)	0.879 ± 0.120	0.849 ± 0.132	0.891 ± 0.015	0.783 ± 0.123	0.855 ± 0.128
LinkNet (SEResNet18) (Chaurasia & Culurciello, 2017)	0.889 ± 0.124	0.855 ± 0.129	0.893 ± 0.120	0.780 ± 0.117	0.869 ± 0.127
FPN (VGG16) (Lin et al., 2017)	0.887 ± 0.122	0.855 ± 0.136	0.894 ± 0.119	**0.803 ± 0.123**	0.861 ± 0.127
FPN (SEResNet18) (Lin et al., 2017)	**0.893 ± 0.119**	0.865 ± 0.129	0.900 ± 0.124	0.796 ± 0.121	**0.877 ± 0.124**
PSPNet (VGG16) (Zhao et al., 2017)	0.880 ± 0.120	0.853 ± 0.138	0.798 ± 0.115	0.752 ± 0.168	0.855 ± 0.130
PSPNet (SEResNet18) (Zhao et al., 2017)	0.868 ± 0.125	0.841 ± 0.133	0.791 ± 0.121	0.723 ± 0.159	0.868 ± 0.130

6.2. Illuminant estimation results

For testing, we used three different traditional methods as well as three learning-based methods that were created for multi-illuminant estimation. Two of the traditional methods employ the Grey-Edge framework (Van De Weijer et al., 2007) as a basis for their approach. This framework contains a large number of methods, and for the sake of simplicity, only the best-performing Grey-Edge method was used for each illumination estimation method. The methods in the framework are very simple. For example, the White-Patch method (Land, 1977) uses the assumption a surface that perfectly reflects light will have the color of the illuminant. To estimate the illuminant, the highest value of each color channel is extracted. The values are combined and normalized to create the RGB illuminant vector. For the learning-based methods, we used three different convolutional neural networks (CNN).

All the traditional methods divide the image into small regions. They assume each region is illuminated by only one illuminant which is estimated for each region. Gijsenij et al. (2012) uses three different approaches to segment the image, uniformly shaped patches, superpixels, and keypoint regions. Hussain and Akbari (2018) divides the image into four regions using Euclidean distance. Beigpour et al. (2013) divides the image into uniformly shaped patches. Beigpour et al. (2013) they use Conditional Random Fields to find the optimal illuminant for each patch from the set of extracted patches.

Bianco et al. (2017) and Shi et al. (2016a) divide the image into patches and perform single-illuminant estimation for each patch. The difference between the models is that Bianco et al. (2017) outputs one prediction per patch and Shi et al. (2016a) uses two neural networks. One network gives two predictions per patch, and the other network selects which prediction is more accurate. Domislović et al. (2021) uses the entire image for multi-illuminant estimation. The model is a modified version of FC4 (Hu et al., 2017). It contains two outputs, one for each illuminant and an attention mechanism for each output. The attention mechanism is used to ignore the regions of the image that contain the wrong illuminant.

Again, the results of the traditional methods are the same in Tables 3 and 4. Unlike learning-based methods, traditional methods need no training, and testing devolves into testing on all images. Unlike the image segmentation problem where the learning-based methods have significantly better performance, the traditional and learning-based methods have comparable results. The best-performing method for the Use-All protocol is the method from Domislović et al. (2021). Another thing that can be observed in Table 3 is the fact that Direct illumination is much easier to estimate than Ambient illumination. This makes sense since Direct illumination is stronger and has a single source, while Ambient illumination is weaker and is the combination of the illumination and the reflections of all the surfaces in the scene.

The results of the One-to-Many protocol can be viewed in Table 4, in which we can recognize a significantly better performance of traditional methods in most metrics. The exceptions are the worst 25% for Ambient and Both and the mean for Ambient, where the method from Domislović et al. (2021) has the best performance. The reason why traditional methods outperform learning methods is that they do not depend on data, and in the case of this protocol, they do not overfit on one camera sensor.

Examining Table 5, we can see that the best performance is obtained on the Outdoor images and the worst on the Indoor images, since Outdoor images contain only natural lighting, Nighttime images contain only artificial lighting, and Indoor images contain both types of lighting. Fig. 4 also shows us that the illumination gamut of Outdoor images is the smallest and is, therefore, easiest to estimate. When looking at Outdoor and Indoor images the best performing method is the method from Domislović et al. (2021) and the best method for Nighttime images is the Patch-based variant of the method introduced in Gijsenij et al. (2012).

7. Conclusion

In this paper, we introduce a novel large-scale multi-illuminant dataset containing 2500 images taken by five different cameras. Each photo is accompanied by a segmentation mask file and an extracted illumination file. The created dataset can be used for both illumination estimation and image segmentation. The dataset follows GDPR privacy regulations, and all sensitive data has been masked. We give a detailed overview of how the images in the dataset were collected and labeled. Moreover, we propose three different evaluation protocols. One of these protocols can be used to evaluate how a method performs when it encounters an image from an unknown camera. Another can be used to see how a method performs with different illuminant and scene types. We tested several methods from the literature on our proposed dataset and compared their results.

Table 3
The mean, median best 25%, and worst 25% angular error scores of different methods tested using the Use-All protocol. Direct represents when only the direct light source illuminant estimation accuracy is examined. Ambient represents when only the ambient light source illumination estimation accuracy is examined. Both represents how well the methods perform in general. The best results are bolded.

Method	Ambient				Direct				Both			
	Mean	med.	Best 25%	Worst 25%	Mean	med.	Best 25%	Worst 25%	Mean	med.	Best 25%	Worst 25%
Hussain and Akbari (2018)	13.19	13.09	6.55	20.09	14.15	13.72	8.72	20.35	13.67	13.46	7.54	20.22
CRF (White-Patch) (Beigpour et al., 2013)	8.40	6.84	1.74	17.96	5.98	4.56	1.36	13.03	7.19	5.44	1.52	15.81
Patch-based (White-Patch) (Gijsenij et al., 2012)	4.89	3.00	0.95	12.22	3.70	2.81	1.09	7.92	4.30	2.89	1.02	10.13
Keypoint-based (White-Patch) (Gijsenij et al., 2012)	6.90	4.52	1.32	16.62	4.01	2.97	0.96	8.91	5.46	3.59	1.11	13.15
Superpixel-based (2nd Order) (Gijsenij et al., 2012)	5.00	3.63	1.26	11.25	3.39	2.71	0.97	7.08	4.20	3.10	1.09	9.32
Bianco et al. (2017)	9.17	7.36	3.43	18.04	6.85	4.58	1.45	16.91	8.01	5.65	2.15	17.98
HypNet/SelNet (Shi et al., 2016b)	6.20	4.20	0.92	14.94	6.41	3.66	0.78	16.90	6.31	3.95	0.85	15.95
Domislović et al. (2021)	**2.84**	**2.13**	**0.74**	**6.21**	**1.71**	**1.22**	**0.44**	**3.86**	**2.28**	**1.60**	**0.55**	**5.22**

Table 4
The mean angular error score of different methods tested using the One-to-Many protocol. Direct represents when only the direct light source illuminant estimation accuracy is examined. Ambient represents when only the ambient light source illumination estimation accuracy is examined. Both represents how well the methods perform in general. The best results are bolded.

Method	Ambient				Direct				Both			
	Mean	med.	Best 25%	Worst 25%	Mean	med.	Best 25%	Worst 25%	Mean	med.	Best 25%	Worst 25%
Hussain and Akbari (2018)	13.19	13.09	6.55	20.09	14.15	13.72	8.72	20.35	13.67	13.46	7.54	20.22
CRF (White-Patch) (Beigpour et al., 2013)	8.40	6.84	1.74	17.96	5.98	4.56	1.36	13.03	7.19	5.44	1.52	15.81
Patch-based (White-Patch) (Gijsenij et al., 2012)	4.89	**3.00**	**0.95**	12.22	**3.70**	2.81	1.09	7.92	4.30	**2.89**	**1.02**	10.13
Keypoint-based (White-Patch) (Gijsenij et al., 2012)	6.90	4.52	1.32	16.62	4.01	2.97	**0.96**	8.91	5.46	3.59	1.11	13.15
Superpixel-based (2nd Order) (Gijsenij et al., 2012)	5.00	3.63	1.26	11.25	3.39	**2.71**	0.97	**7.08**	**4.20**	3.10	1.09	9.32
Bianco et al. (2017)	10.09	8.21	3.17	20.40	9.48	7.39	3.26	20.02	9.78	7.67	3.21	20.36
HypNet/SelNet (Shi et al., 2016b)	8.35	6.00	1.31	19.41	7.67	4.98	1.05	19.12	8.01	5.45	1.17	19.31
Domislović et al. (2021)	**4.83**	4.18	1.83	**9.05**	4.58	4.20	1.89	7.85	4.71	4.19	1.86	**8.45**

Table 5
The mean, median best 25%, and worst 25% angular error scores of different methods tested on the three variants of the By-Type protocol. Direct represents when only the direct light source illuminant estimation accuracy is examined. Amb. represents when only the ambient light source illumination estimation accuracy is examined. Both represents how well the methods perform in general. The best results are bolded.

Method	Outdoor			Indoor			Nighttime		
	Amb.	Direct	Both	Amb.	Direct	Both	Amb.	Direct	Both
Hussain and Akbari (2018)	13.44	12.94	13.19	15.42	14.26	14.84	16.91	15.79	16.35
CRF (White-Patch) (Beigpour et al., 2013)	9.00	5.62	7.31	8.16	9.40	8.78	7.10	6.81	6.96
Patch-based (2nd Order) (Gijsenij et al., 2012)	5.21	3.84	4.53	5.76	5.40	5.58	**4.39**	**2.57**	**3.48**
Keypoint-based (White-Patch) (Gijsenij et al., 2012)	7.39	4.15	5.77	6.99	5.50	6.25	5.06	2.93	3.99
Superpixel-based (White-Patch) (Gijsenij et al., 2012)	5.88	3.59	4.73	5.78	4.57	5.18	5.98	3.29	4.63
Bianco et al. (2017)	3.74	5.98	4.86	7.98	6.85	7.42	7.61	8.82	8.22
HypNet/SelNet (Shi et al., 2016b)	5.99	6.26	6.09	6.49	7.65	7.07	5.28	4.24	4.76
Domislović et al. (2021)	**2.26**	**1.30**	**1.78**	**4.72**	**4.18**	**4.45**	4.42	2.93	3.68

CRediT authorship contribution statement

Ilija Domislović: Conceptualization, Methodology, Writing – original draft, Writing – review & editing. **Donik Vršnak:** Validation, Software, Data curation. **Marko Subašić:** Writing – review & editing, Supervision, Formal analysis. **Sven Lončarić:** Writing – review & editing, Supervision, Project administration.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The dataset is available at http://bit.ly/shal_dataset.

References

Banić, N., Koščević, K., & Lončarić, S. (2017). Unsupervised learning for color constancy. arXiv preprint arXiv:1712.00436.

Beigpour, S., Riess, C., Van De Weijer, J., & Angelopoulou, E. (2013). Multi-illuminant estimation with conditional random fields. *IEEE Transactions on Image Processing*, *23*(1), 83–96. http://dx.doi.org/10.1109/CVPR42600.2020.00332.

Bianco, S., Cusano, C., & Schettini, R. (2017). Single and multiple illuminant estimation using convolutional neural networks. *IEEE Transactions on Image Processing*, *26*(9), 4347–4362. http://dx.doi.org/10.1109/TIP.2017.2713044.

Bleier, M., Riess, C., Beigpour, S., Eibenberger, E., Angelopoulou, E., Tröger, T., & Kaup, A. (2011). Color constancy and non-uniform illumination: Can existing algorithms work? In *2011 IEEE international conference on computer vision workshops (ICCV Workshops)* (pp. 774–781). IEEE, http://dx.doi.org/10.1109/ICCVW.2011.6130331.

Chaurasia, A., & Culurciello, E. (2017). LinkNet: Exploiting encoder representations for efficient semantic segmentation. In *2017 IEEE visual communications and image processing* (pp. 1–4). http://dx.doi.org/10.1109/VCIP.2017.8305148.

Cheng, D., Prasad, D. K., & Brown, M. S. (2014). Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. *Journal of the Optical Society of America A*, *31*(5), 1049–1058. http://dx.doi.org/10.1364/JOSAA.31.001049, URL: http://opg.optica.org/josaa/abstract.cfm?URI=josaa-31-5-1049.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255). Ieee.

Domislović, I., Vrsnak, D., Subašić, M., & Lončarić, S. (2021). Outdoor daytime multi - illuminant color constancy. In *2021 12th international symposium on image and signal processing and analysis* (pp. 270–275). http://dx.doi.org/10.1109/ISPA52656.2021.9552092.

Fairchild, M. D. (2013). *Color appearance models*. John Wiley & Sons, Ltd, http://dx.doi.org/10.1002/9781118653128, URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118653128, arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781118653128.

Finlayson, G. D., Drew, M. S., & Funt, B. V. (1993). Diagonal transforms suffice for color constancy. In *1993 (4th) international conference on computer vision* (pp. 164–171). IEEE, http://dx.doi.org/10.1109/ICCV.1993.378223.

Finlayson, G. D., Hemrit, G., Gijsenij, A., & Gehler, P. (2017). A curious problem with using the colour checker dataset for illuminant estimation. In *Color and imaging conference, Vol. 25* (pp. 64–69). Society for Imaging Science and Technology.

Gehler, P. V., Rother, C., Blake, A., Minka, T., & Sharp, T. (2008). Bayesian color constancy revisited. In *2008 IEEE conference on computer vision and pattern recognition* (pp. 1–8). IEEE, http://dx.doi.org/10.1109/CVPR.2008.4587765.

Gijsenij, A., Gevers, T., & Van De Weijer, J. (2011). Computational color constancy: Survey and experiments. *IEEE Transactions on Image Processing*, *20*(9), 2475–2489. http://dx.doi.org/10.1109/TIP.2011.2118224.

Gijsenij, A., Lu, R., & Gevers, T. (2011). Color constancy for multiple light sources. *IEEE Transactions on Image Processing*, *21*(2), 697–707. http://dx.doi.org/10.1109/TIP.2011.2165219.

Gijsenij, A., Lu, R., & Gevers, T. (2012). Color constancy for multiple light sources. *IEEE Transactions on Image Processing*, *21*(2), 697–707. http://dx.doi.org/10.1109/TIP.2011.2165219.

Hordley, S. D. (2006). *Scene illuminant estimation: past, present, and future, Vol. 31* (pp. 303–314). Color Research & Application: Endorsed By Inter-Society Color Council, the Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, the Swedish Colour Centre Foundation, Colour Society of Australia, Centre FranÇAis de la Couleur, http://dx.doi.org/10.1002/col.20226.

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *2018 IEEE/CVF conference on computer vision and pattern recognition* (pp. 7132–7141). http://dx.doi.org/10.1109/CVPR.2018.00745.

Hu, Y., Wang, B., & Lin, S. (2017). Fc4: Fully convolutional color constancy with confidence-weighted pooling. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 4085–4094). http://dx.doi.org/10.1109/CVPR.2017.43.

Hussain, M. A., & Akbari, A. S. (2018). Color constancy algorithm for mixed-illuminant scene images. *IEEE Access*, *6*, 8964–8976. http://dx.doi.org/10.1109/ACCESS.2018.2808502.

Laakom, F., Raitoharju, J., Nikkanen, J., Iosifidis, A., & Gabbouj, M. (2021). Intel-tau: A color constancy dataset. *IEEE Access*, *9*, 39560–39567. http://dx.doi.org/10.1109/ACCESS.2021.3064382.

Land, E. H. (1977). The retinex theory of color vision. *Scientific American*, *237*(6), 108–129.

Leung, C.-K., & Lam, F.-K. (1994). Image segmentation using maximum entropy method. In *Proceedings of ICSIPNN '94. international conference on speech, image processing and neural networks, Vol. 1* (pp. 29–32). http://dx.doi.org/10.1109/SIPNN.1994.344973.

Lin, T.-Y., Dollar, P., Girshick, R., He, K., Hariharan, B., & Belongie, S. (2017). Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 2117–2125). http://dx.doi.org/10.1109/CVPR.2017.106.

Otsu, N. (1979). A threshold selection method from Gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, *9*(1), 62–66. http://dx.doi.org/10.1109/TSMC.1979.4310076.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In N. Navab, J. Hornegger, W. M. Wells, & A. F. Frangi (Eds.), *Medical image computing and computer-assisted intervention – MICCAI 2015* (pp. 234–241). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-24574-4_28.

Shi, W., Loy, C. C., & Tang, X. (2016a). Deep specialized network for illuminant estimation. In *European conference on computer vision* (pp. 371–387). Springer.

Shi, W., Loy, C. C., & Tang, X. (2016b). Deep specialized network for illuminant estimation. In B. Leibe, J. Matas, N. Sebe, & M. Welling (Eds.), *Computer vision – ECCV 2016* (pp. 371–387). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-319-46493-0_23.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. URL: http://arxiv.org/abs/1409.1556 arXiv:1409.1556 [cs].

Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. In V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, & I. Maglogiannis (Eds.), *Lecture notes in computer science*, *Artificial neural networks and machine learning – ICANN 2018* (pp. 270–279). Cham: Springer International Publishing, http://dx.doi.org/10.1007/978-3-030-01424-7_27.

Van De Weijer, J., Gevers, T., & Gijsenij, A. (2007). Edge-based color constancy. *IEEE Transactions on Image Processing*, *16*(9), 2207–2214. http://dx.doi.org/10.1109/TIP.2007.901808.

von Kries, J. (1905). Influence of adaptation on the effects produced by luminous stimuli. *Handbuch der Physiologie des Menschen*, *3*, 109–282, URL: https://ci.nii.ac.jp/naid/10030415665/en/.

Wang, K., Chen, Z., Wu, Q. M. J., & Liu, C. (2019). Face recognition using AMVP and WSRC under variable illumination and pose. *Neural Computing and Applications*, *31*(8), 3805–3818. http://dx.doi.org/10.1007/s00521-017-3316-x.

Xiao, J., Gu, S., & Zhang, L. (2020). Multi-domain learning for accurate and few-shot color constancy. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3258–3267). http://dx.doi.org/10.1109/CVPR42600.2020.00332.

Zhao, H., Shi, J., Qi, X., Wang, X., & Jia, J. (2017). Pyramid scene parsing network. In *2017 IEEE conference on computer vision and pattern recognition* (pp. 6230–6239). http://dx.doi.org/10.1109/CVPR.2017.660.