IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 16 December 2022, accepted 31 December 2022, date of publication 4 January 2023, date of current version 9 January 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3234115

RESEARCH ARTICLE

Framework for Illumination Estimation and Segmentation in Multi-Illuminant Scenes

DONIK VRŠNAK, (Graduate Student Member, IEEE), ILIJA DOMISLOVIĆ, (Graduate Student Member, IEEE), MARKO SUBAŠIĆ, (Member, IEEE), AND SVEN LONČARIĆ, (Senior Member, IEEE)

Image Processing Laboratory, Faculty of Electrical Engineering and Computing, University of Zagreb, 10000 Zagreb, Croatia

Corresponding author: Donik Vršnak (donik.vrsnak@fer.hr)

ABSTRACT Color constancy is an important part of the human visual system, as it allows us to perceive the colors of objects invariant to the color of the illumination that is illuminating them. Modern digital cameras have to be able to recreate this property computationally. However, this is not a simple task, as the response of each pixel on the camera sensor is the product of the combination of spectral characteristics of the illumination, object, and the sensor. Therefore, many assumptions have to be made to approximately solve this problem. One common procedure was to assume only one global source of illumination. However, this assumption is often broken in real-world scenes. Thus, multi-illuminant estimation and segmentation is still a mostly unsolved problem. In this paper, we address this problem by proposing a novel framework capable of estimating per-pixel illumination of any scene with two sources of illumination. The framework consists of a deep-learning model capable of segmenting an image into regions with uniform illumination and models capable of single-illuminant estimation. First, a global estimation of the illumination is produced, and is used as input to the segmentation model along with the original image, which segments the image into regions where that illuminant is dominant. The output of the segmentation is used to mask the input and the masked images are given to the estimation models, which produce the final estimation of the illuminations. The models comprising the framework are first trained separately, then combined and fine-tuned jointly. This allows us to utilize well researched single-illuminant estimation models in a multi-illuminant scenario. We show that such an approach improves both segmentation and estimation capabilities. We tested different configurations of the proposed framework against other single- and multi-illuminant estimation and segmentation models on a large dataset of multi-illuminant images. On this dataset, the proposed framework achieves the best results, in both multi-illumination estimation and segmentation problems. Furthermore, generalization properties of the framework were tested on often used single-illuminant datasets. There, it achieved comparable performance with state-of-the-art single-illumination models, even though it was trained only on the multi-illuminant images.

INDEX TERMS Color constancy, segmentation, multi-illuminant, illumination estimation, deep learning, framework.

I. INTRODUCTION

Color constancy is an important part of the human visual system, as it allows us to adapt to different colors of illumination. This enables us to recognize the colors of objects and illuminants independently. For images taken by digital cameras, it is essential to be able to estimate the color of illumination as accurately as possible. Accurate estimation allows us to create a fateful reproduction of the scene which is satisfactory to the human observer. Furthermore, inaccurate estimation creates images that are influenced by illumination, which can decrease the performance of downstream image processing tasks, as described in [1]. Thus, computational color constancy has been studied by numerous authors since

The associate editor coordinating the review of this manuscript and approving it for publication was Essam A. Rashed.

the advent of digital cameras, and many methods have been proposed. Equation (1) describes the amount of light $p_c(x, y)$ recorded for each channel $c \in R, G, B$ at the position x, y in the scene:

$$p_c(x, y) = \int_\omega I(x, y, \lambda) R(x, y, \lambda) S_c(\lambda) d\lambda \,, \tag{1}$$

where $R(\lambda)$ and $I(\lambda)$ are the reflectivity and illumination spectral functions, respectively. S_c represents the spectral sensitivity of the observer (camera). Equation (1) also shows that for each value p_c, there are an infinite number of combinations of I, R, S that can produce it. Since I, R, S are normally unknown, that makes the problem of illumination estimation under constrained.

However, assumptions about the properties of the illumination or the properties of the scene can be introduced. This makes it feasible to approximate the value of the illumination present in the scene. That step of computational color constancy is called illuminant estimation. The second step is to white-balance the image, usually to make it look as it was taken under a canonical illuminant, thus eliminating the influence of the illumination. The most common approximation used is the von Kries [2] model:

$$\begin{bmatrix} c_R \\ c_B \\ c_G \end{bmatrix} = \begin{bmatrix} e_{c,R}/e_{u,R} & 0 & 0 \\ 0 & e_{c,B}/e_{u,B} & 0 \\ 0 & 0 & e_{c,G}/e_{u,G} \end{bmatrix} \begin{bmatrix} p_R \\ p_B \\ p_G \end{bmatrix} \tag{2}$$

where $[c_R \;\; c_B \;\; c_G]^T$ represents the corrected image, and $[p_R \; p_B \; p_G]^T$ is the value retrieved from the sensor. Canonical illumination is represented by $\boldsymbol{e}_c$ and $\boldsymbol{e}_u$ is the estimated illuminant. While Equation (2) does not provide true compensation for the illumination, it is an approximation that works well.

Different assumptions have been applied to the problem of illuminant estimation. One such assumption is that there was only one illuminant present in the scene. However, for many real-world scenes that is not the case. They contain at least two sources of illumination, e.g., outdoor scenes that are illuminated with direct sunlight and with shaded areas illuminated by skylight, or indoor scenes where one illuminant is a light bulb and the other is the sunlight coming through the window. For such scenes, illuminant localization is as important as the estimation, as just the color of the illumination does not provide enough information for accurate correction of the image. Figure 1 shows an example of a real-world scene with two illuminants. The effects of global correction are also shown.

In this work, we propose a novel deep learning framework that is capable of both segmentation and estimation of scenes with two sources of illumination. The main idea behind the framework is to separate the problem of illuminant localization and estimation to different specialized methods. This allows us to utilize well-researched single-illuminant estimation models for multi-illuminant scenes. The framework is composed of three main steps. First, a global illumination vector for the image is estimated. Next, this illumination vector is fed into a segmentation model, alongside the original image. The output of this step is the segmentation mask, which shows where the first estimated illuminant is dominant in the scene. Then, the original image is masked, and the masked images are fed to global estimation models. The outputs of estimation models are combined with the segmentation mask to produce the final estimation of the illumination in the whole scene.

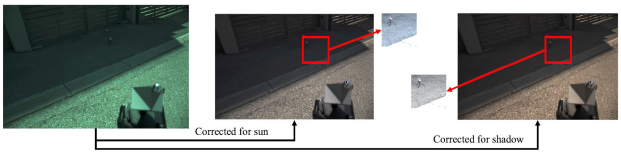

FIGURE 1. Two corrections by different illuminants present in the raw image, with gamma correction applied for easier visualization. The middle image is corrected for the sunlight, and the shaded regions end up having a blue hue. The right image is corrected for illumination in the shadow, which corresponds to the blueish skylight. This gives the sunlit region an orange cast. Groundtruth values were obtained from the gray sides of the SpyderCube calibration object that are highlighted by red squares.

Furthermore, we incorporate the possibility of illuminant mixing, and the proposed framework is capable of providing a per-pixel estimation of illumination. This is achieved by linear combination of estimated illuminants using the segmentation mask. We show that by incorporating joint end-to-end training of the framework, we achieve state-of-the-art results. Additionally, we show that joint training further improves the performance of underlying models when compared to the same models that were only trained separately. The training of the framework was done on a large multi-illuminant dataset [3] containing 2500 indoor and outdoor images. Testing was done on a hold-out set of images from the dataset for the multi-illuminant scenario. The generalization performance was tested by training the framework on multi-illuminant images and testing it on single-illuminant images from Cube+ [4] and ColorChecker [5] datasets. On multi-illuminant images, the proposed framework achieves state-of-the-art results. It also achieves results comparable with best single-illuminant estimation methods on single-illuminant datasets. Furthermore, usage of separate models for each sub-task makes the framework modular. This allows us to easily train and test different variations of the framework, thus balancing the accuracy with complexity. We describe our framework in detail in Section III, and show the quantitative and qualitative results for both the segmentation and per-pixel estimation compared to other single- and multi-illuminant methods in Section IV. Finally, in Section V we conclude the paper.

II. RELATED WORK

The term computational color constancy usually includes two basic steps. These are illumination estimation and color correction (also referred to as chromatic adaptation). The first step is determining the illumination vector for some part

of the image. The granularity of the estimation can vary, from per-pixel, through image patches all the way up to the whole image. This defines the type of estimation method that is needed, with single-illuminant estimation methods estimating only one illuminant for the whole input image. Patch and per-pixel estimations fall under multi-illumination estimation methods, as they estimate more than one illuminant per image. Since color constancy is an ill-posed problem, most research in the past focused on the problem of single-illuminant color constancy. With this assumption, it is assumed that the whole scene (or at least the vast majority) is illuminated by one global illuminant. One of the first methods for single-illuminant color constancy methods were simple methods that relied on low-level image statistics. Two of those methods are the Gray-World [6] and the White-Patch [7] (Max RGB) methods. Gray-World method assumes that for each scene, the average reflectance under white light is gray, and thus any deviation from gray is caused by the color of the illumination. On the other hand, the White-Patch (Max-RGB) method assumes that the brightest part of the scene is the reflected color of the illuminant from a specular surface. However, it is easy to find common real-world examples where these assumptions are broken. For example, for the Gray-World method, any scene with numerous plants (like forests and parks) will not have a gray reflectance under white light but instead that average will be green. For White-Patch, if the scene does not contain any specular highlights, the assumption will be broken. More complex methods were proposed over the years. They can be split into two main categories, statistics based and learning-based methods. Some of the more well-known statistics-based methods include the Gray-Edge framework [8], which generalizes all methods such as Gray-World and White-Patch by adding the possibility of using image gradients and different image norms, as described by (3):

$$\left(\int \left| \frac{\partial^n f_{c,\sigma}(x)}{\partial x^n} \right|^p dx \right)^{\frac{1}{p}} = k e_c^{n,p,\sigma}, \tag{3}$$

where $|\cdot|$ is the Frobenius norm, $c \in R, G, B$, n is the order of the derivative and p is the Minkowski-norm.

There are also gamut mapping methods, such as the method proposed in [9]. There, the goal of the method is to find the gamut that the illuminant spans in the chromaticity diagram and then use that knowledge to find the most probable illuminant color. On the other hand, learning-based methods are more complex, and can be split into two categories: simpler machine learning methods and more complex deep learning methods. One of these learning-based methods [10] learns the common surfaces in the train scenes and then uses the exemplar approach to match the surfaces in the test images to those learned surfaces. Other methods, such as [11], [12], and [13] use a probabilistic model of the illumination and reflectance as a random variable. Unfortunately, all of these methods do not achieve good enough results, particularly in more challenging conditions.

This is the reason more complex deep learning models were proposed for the task of color constancy. The first attempt at such a model was proposed in [14], where a simple network was given a raw image and produced the estimation of the illumination in the scene. Because there was no large dataset, this method was trained mostly on image patches. However, this reduced the semantic information present in each patch, and eliminated cross patch information. This was addressed in [15], where the authors proposed a method that took as input the whole image and produced estimation for patches of the image. Additionally, the method produced an attention map which was used to multiply the patch estimates and produce the final estimation mask. This approach was successful because it allowed the model to reason about the patches of the image that carry more information about the color of the illumination. In [16], the authors propose a very deep model for illuminant estimation (CRNA) that uses cascading residual connections and ResNet architecture to stabilize learning and improve performance. Similarly, in [17], the authors propose a deep network which iteratively estimates the illumination, which is also used to stabilize training and improve performance. On the other hand, in [18], a small network that still achieves state-of-the-art results for illuminant estimation is proposed. Furthermore, some methods, such as [19] and [20], use only image histograms with the deep learning models to perform illuminant estimation. This removes any spatial information and focuses only on colors present in the scene.

On the other hand, multi-illuminant color constancy has been much less studied in the past than single-illuminant color constancy. One reason for this is the lack of a large multi-illuminant dataset, since it is difficult to accurately annotate multi-illuminant images. Most of the methods that were proposed for this problem are learning-based and model the spatial distribution of illuminants. However, several statistics-based methods have been proposed in [21], [22], [23], and [24]. They share some similarity with our approach, as they separate segmentation and estimation into separate tasks that are combined. They use image texture [23] or Kmeans [22] for localization and then use Max RGB method for estimation. Finally, similar to our method, the localization is used to compute the final per-pixel illumination of the scene. On the other hand, [25] propose a white-balancing method for scenes in which the total number of illuminants is not known. They achieve this by selecting N white-balance points and map them to ground truth ones. Finally, [26] proposed a method that imitates the Adaptive Surround Modulation (ASM) capability of the human eye to regulate the receptive field of neurons based on contrast. One classical machine learning approach was presented in [27], where the authors use conditional random fields to create the MIRF algorithm, which can localize and estimate illuminants in the scene. The main drawback of this approach is its high computational cost and lower accuracy. Deep learning-based approach for multi-illuminant color constancy was proposed in [28] as an upgrade on the network proposed in [14], where

the authors use kernel density estimation to determine the number of illuminants in the scene. In [29], authors propose a framework of two networks, HypNet and SelNet. HypNet network proposes two hypotheses about the illumination of each patch, and SelNet chooses which of those hypotheses to use for the estimation.

More recently, in [30] the authors proposed a simple model that used brightness threshold to perform image segmentation, to which they applied simple estimation methods. This method works very fast, but it produces many artifacts and incorrect corrections in parts of the scene where the brightness assumption does not hold. Furthermore, three methods for image segmentation and estimation using deep learning models were proposed. In [31], the authors introduced a vision transformer method that was able to perform segmentation of parts of the scene that were incorrectly white-balanced. In [32], the authors proposed an autoencoder training strategy and a novel loss function which was capable of learning the common distribution of colors in scenes, to produce per-pixel estimation of the illumination. Finally, in [33] the authors created a segmentation model that was able to segment scenes with two sources of illumination by first producing an estimation of the primary illuminant. We based our framework on the same principle: that it is possible to relatively accurately estimate one of the illuminant sources in the scene using global methods, and then localize its influence. However, unlike the model in [33] we do not stop at segmentation, as our framework allows for accurate estimation of both illuminant sources and their localization.

III. PROPOSED FRAMEWORK

In this work, we present a novel framework for simultaneous estimation and segmentation of illumination for scenes with two sources of illumination. The main idea behind our framework was to leverage well-researched single-illuminant estimation models for multi-illuminant scenes. The proposed framework consists of three main parts. A scheme of the framework can be found in Figure 2. The first part is global estimation of the dominant illuminant. Then, a segmentation model is used to localize the influence of the dominant illuminant, which is represented as a binary segmentation mask. This mask is used to create masked inputs for the two estimation models. Then, those two single-illuminant estimation models are used to estimate the dominant and secondary illuminants. Finally, the per-pixel estimation of illumination for the scene is obtained by linear combination of the estimated illuminants based on the weights from the segmentation output, using Equation (4):

$$p(x, y)_c = (1 - S_p(x, y))I_{\mathrm{p1}} + S_p(x, y)I_{\mathrm{p2}}\,, \tag{4}$$

where (x, y) are the coordinates in the image, p_c is the final per-pixel estimation, I_{p1} and I_{p2} are dominant and secondary illumination estimations, respectively, and S_p is the predicted segmentation mask.

Each layer of the framework is implemented so that it allows for the free flow of gradients using backpropagation. This allows us to train the framework end-to-end. We refer to this as joint training. Gradients in the upper layers during training of the framework are computed from both the estimation and the segmentation errors. This effect is not present when layers are only trained separately. Another benefit of this approach is in the transitional regions between the illuminations. In those regions, the segmentation model is encouraged to keep the output such that the linear combination of the illumination sources corresponds to the real mixed illumination. Thus, the segmentation output is pushed closer to 0.5 than to 0 or 1 for those areas. In the case of the pure segmentation training, where the goal is to create hard borders between classes, no such regularization effect is present. Furthermore, those regions carry less useful information for either of the single-illuminant estimation models that come after the segmentation. For them, this ambiguity in the segmentation acts as an attention mechanism, by shifting focus more to the parts of the scene where illumination is less ambiguous. We show later that this type of joint training of our framework improves the performance of both segmentation and estimation model compared with their counterparts that were trained independently.

Moreover, we propose an additional recurrent component because it can sometimes be difficult to estimate the dominant illuminant from the whole image in the first step. The recurrent connection is shown with a labeled dotted arrow in the red part of Figure 2. It naturally follows that, if we can localize and estimate one illuminant in the scene, the estimation produced would be better than the global estimation. Thus, the recurrent component enables additional passes through the framework. In the second pass through the framework, the recurrent connection replaces the initial global dominant illuminant estimation with the output of the local dominant illuminant estimation from the first pass. For the final output of the framework, all the intermediary estimation and segmentation steps are averaged. Such recurrent behavior can be implemented in as many steps as it is necessary. However, since the task of color constancy usually needs to be performed quickly, we implemented only a two-step recurrent framework. We compare the performance of this recurrent framework to that of the base framework as well as other multi- and single-illuminant models in Section IV.

For the estimation task, the framework is designed in such a way that it is interoperable with any state-of-the-art single-illuminant estimation methods. In the scope of this paper, we implemented a single-illuminant estimation model based on the FC4 [15] model, with a reduced number of parameters. We reduced the number of parameters to decrease the overall complexity of the framework. We use one of these models to first predict the dominant illuminant in the scene. Later, we use two more such models to predict the illumination in the regions highlighted by the segmentation model. Furthermore, in some variations of our framework, the weights are shared between these two models. (In practice, this is implemented with only one estimation model, to reduce memory

usage.) These estimation models are shown in yellow in Figure 2.

Finally, we limit the number of illuminants for two main reasons. Firstly, we are limited by the types of datasets that are available for multi-illuminant scenes, which are needed to train our model. All the labeled datasets that have per-pixel groundtruth information about the illuminants in the scene contain only two illuminants. Moreover, most real-world scenes actually contain either one or two illuminants. One exception are very dynamic nighttime scenes like clubs or urban areas. However, we show that even with this reduction in the number of illuminants, our model can handle complex scenes. We achieve this by allowing illuminant mixing, which is very common in real-world scenes. Furthermore, the results show that our model preforms well on single-illuminant scenes, even though it was trained only on scenes with two sources of illumination.

A. TRAINING

The framework was trained in two steps. First, each component of the framework was trained on their respective task separately. The segmentation part was trained to segment the areas of the scene where the primary illuminant was dominant, similar to the method proposed in [33]. The estimation models were trained to predict either the dominant or the secondary illuminant. After the pretraining step, the framework was combined into the final model as described in Section III and then trained end to end using backpropagation. The framework was implemented in TensorFlow 2.4 and trained on a system with an RTX 2080Ti GPU and AMD Ryzen 3700x CPU. Pretraining was done over 500 epochs, with cosine annealing scheduler [34] and stochastic gradient descent [35] optimizer. We use a linear combination of the binary cross entropy (BCE) and robust color constancy loss (IL) function [36] for the segmentation and estimation outputs, respectively. This combined loss can be expressed as:

$$\begin{aligned} &\mathrm{L}(I_{\mathrm{p0}}, I_{\mathrm{p1}}, I_{\mathrm{p2}}, S_p, I_{\mathrm{gt1}}, I_{\mathrm{gt2}}, S_{\mathrm{gt}}) \\ &\quad = \alpha \mathrm{IL}(I_{\mathrm{p0}}, I_{\mathrm{gt1}}) \\ &\qquad + \beta \mathrm{BCE}(S_p, S_{\mathrm{gt}}) + \gamma \mathrm{IL}(I_{\mathrm{p1}}, I_{\mathrm{gt1}}) + \delta \mathrm{IL}(I_{\mathrm{p2}}, I_{\mathrm{gt2}}) \quad (5) \\ &\mathrm{BCE}(S_p, S_{\mathrm{gt}}) \\ &\quad = -S_{gt} \log(S_p) - (1 - S_{\mathrm{gt}}) \log(1 - S_p) \quad (6) \\ &\mathrm{IL}(I_p, I_{\mathrm{gt}}) \\ &\quad = \left\| \frac{I_p - I_{\mathrm{gt}}}{I_{\mathrm{gt}}} \right\|_2, \quad (7) \end{aligned}$$

where I_{p0} is the initial estimation of the dominant illuminant, I_{p1} and I_{p2} are the final estimations of the dominant and secondary illuminant, and S_p is the predicted segmentation mask. I_{gt1}, I_{gt2}, and S_{gt} are the groundtruth information about the illuminants and the segmentation mask, respectively. BCE (Equation (6)) is the binary cross entropy function applied at the pixel level. The IL (Equation (7)) loss function is the robust color constancy loss function proposed in [36]. Coefficients α, β, γ, and δ were selected using random search of the hyperparameter space and their values were set to 0.7, 1.0, 0.9, and 0.9 respectively.

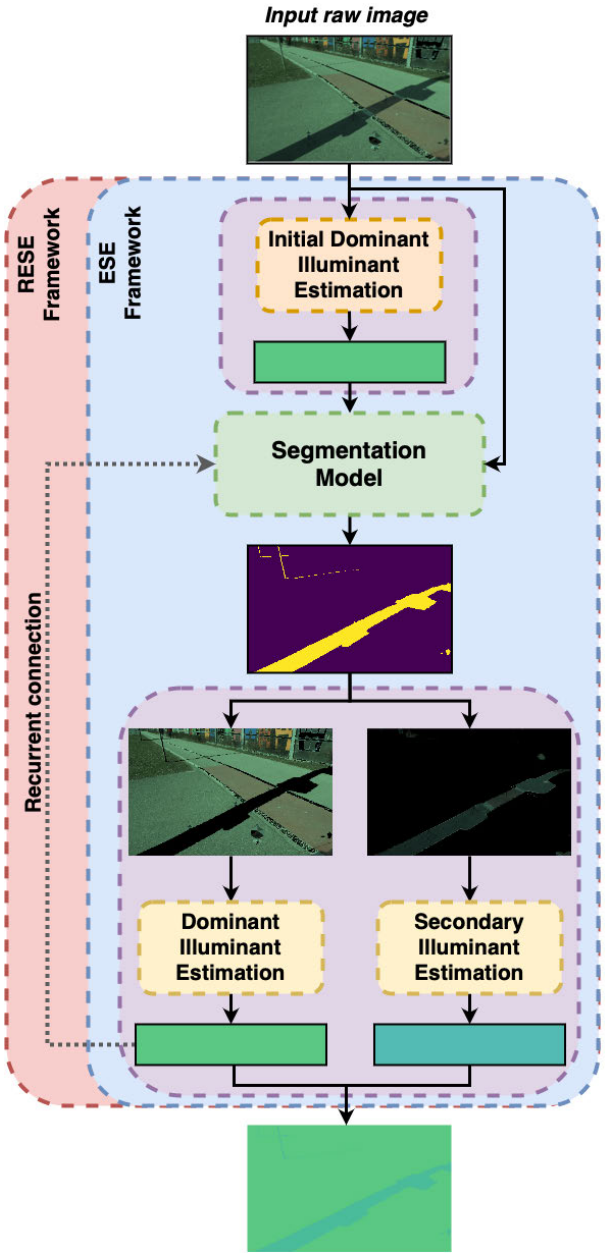

FIGURE 2. Scheme of the proposed framework. In general, the framework consists of the initial estimator of the dominant illuminant, followed by a segmentation model that is capable of localizing the presence of that illuminant in the scene. Then that output is used to create two masked images, which are then given to the estimation models. The estimation models then produce two estimations that are combined to create the final per-pixel estimation of the illumination. The estimation models in the bottom purple box can either be independent or have shared weights. The recurrent extension to our framework is shown in red. The dotted line represents the recurrent connection that allows us to use the dominant illuminant estimation as the input to the segmentation model in the second pass.

Joint training of the framework was done using the same scheduler and optimizer for another 500 epochs. To provide a fair comparison, models that were not trained jointly were all trained for 1000 epochs to eliminate any problems with

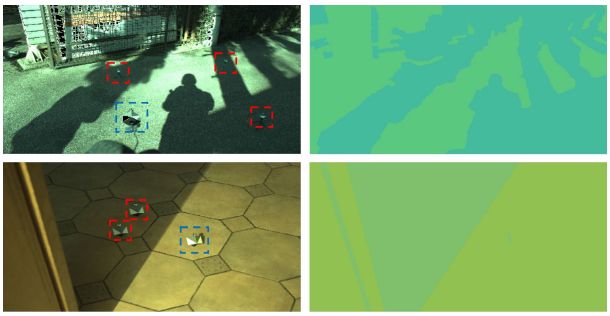

FIGURE 3. Example of the images used for training. The first and second rows show an outdoor and indoor scene, with 4 and 3 SpyderCube calibration objects, respectively. Multiple cubes marked by red squares are placed in the region illuminated by the ambient illuminant, which can vary throughout the scene more than the direct illumination (e.g., sunlight or one light bulb). We used only images where the difference in the ground truth between the measured ambient illumination was less than 1 degree to ensure that the manual annotation of the regions shown in the second column is accurate. Our annotation procedure is similar to that described in [24].

under fitting. The parameters of the model that scored the best on the validation set were taken for testing to prevent overfitting. For the training, we used a newly constructed dataset [3] containing 2500 outdoor and indoor scenes with two sources of illumination, taken by 5 different cameras. All images were manually annotated to contain per-pixel groundtruth illumination values. Few examples of images and the groundtruth from this dataset are shown in Figure 3.

B. PERFORMANCE MEASURES

We compare the model performance on a hold-out set of the two-illuminant dataset on both the segmentation and illumination tasks. For the single-illuminant datasets, we compare only the performance of illuminant estimation. To quantitatively compare the results, we use two metrics, Dice coefficient [37] for segmentation and angular distance for illuminant estimation.

Dice coefficient [37] is computed as:

$$\text{Dice} = \frac{2|\text{TP}|}{2|\text{TP}| + |\text{FP}| + |\text{FN}|}, \tag{8}$$

where TP, FP, FN are true positive, true negative and false negative values when comparing the prediction to the groundtruth. $|\cdot|$ represents the cardinality (number of elements) of the set.

For the estimation task, we use angular error, which can be computed as:

$$\text{err}_{\text{ang}} = \cos^{-1}\left(\frac{\boldsymbol{e}_r \cdot \boldsymbol{e}_p}{\boldsymbol{e}_r \boldsymbol{e}_p}\right), \tag{9}$$

where $\cdot$ denotes vector dot product, $\boldsymbol{e}_r$ is the real illuminant and $\boldsymbol{e}_p$ the estimated illuminant. Since the groundtruth and estimation are pixel-based, we report the average error over the whole image. The classes in the segmentation masks are relatively well-balanced, so the average value of the error is not biased towards either illuminant. For single-illuminant comparison, our model was only trained on the multi-illuminant images, and then tested on the images from the single-illuminant dataset. In this case, we obtain the single-illuminant estimate by applying global average pooling to the per-pixel illuminant estimations.

IV. RESULTS

The models were tested on a hold out set of our dataset [3], and on single-illuminant images from the Cube+ [4] and ColorChecker [5] datasets. Thus, we test the performance of our model in both single- and multi-illuminant scenarios. In the case of the single-illuminant images, the models were trained only on the images from our two-illuminant dataset, and then tested as is on the single-illuminant images. The framework was compared to other methods for both multi-illuminant segmentation and estimation tasks, and these results are shown in Tables 1 and 2 respectively. The comparison of results on single-illuminant images are shown in Tables 3 and 4.

Table 1 shows the results of the segmentation task. The first block of models are the simple baseline models, the second block is the segmentation models implemented from other works. The third block presents the variations of the proposed framework. They show that our framework outperform all other implemented models, and by a solid margin, independent of the number of parameters. The models that were used for comparison include the illumination segmentation models proposed in [32], [33], and [30], U-Net [38] models with VGG-16 and VGG-19 [39] encoders (implemented such that one illuminant was known, as described in [33]) and a baseline Otsu threshold applied to the brightness histogram of the image. It is important to note that the framework performs better than the pure segmentation models (VGG-16). This holds even when the number of parameters is comparable (approx. 34 million parameters in the case of the VGG-19 based autoencoder and VGG-16 based framework). This indicates that the joint training that was used to train our framework increases both the segmentation and estimation parts of our model. To further test this, we compare the jointly trained framework to one whose components were trained only separately (i.e., no joint training was done). Again, we see the improvement in performance, thus providing further evidence of the benefit of joint training (seen in the last block in Table 1). We denote the frameworks where the parameters of the estimation models are shared by omitting the "x2" modifier in the name. RESE denotes the recurrent variant of our framework with two steps.

Since our framework is primarily designed to produce a per-pixel estimate of the illumination, the main focus will be on those results. Table 2 shows the estimation results on our dataset with two illuminants for many multi- and single-illuminant methods that were implemented. In it, the first block of models are the simple baseline models single- and multi-illuminant estimation models. The second block contains the estimation models implemented from other works. The third block contains variations of the framework that were not jointly trained. Finally, the fourth block contains variations of the proposed framework with joint training.

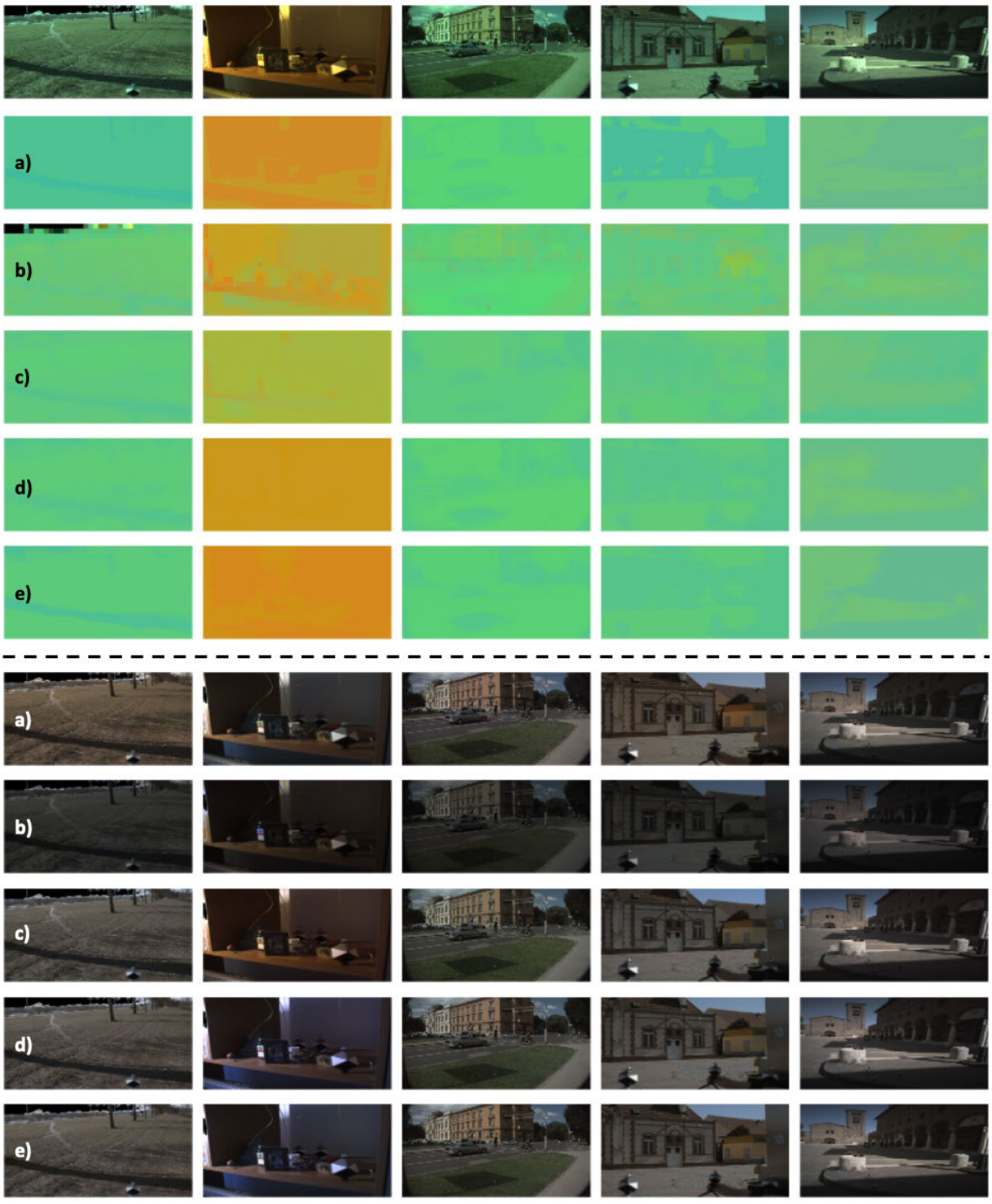

FIGURE 4. Example of the randomly selected images, corresponding groundtruths, estimations and corrections for multi-illuminant estimation methods. The first row is the input image, the first set of images are the per pixel illumination estimations, while the second set are the corrections. In each section, rows are marked with letters corresponding to different models and groundtruth. These are: (a) Groundtruth, (b) Bianco-CNN [28], (c) Autoencoder-based [32] (VGG-16), (d) VGG-16 + FC4 × 2 (non jointly trained framework), and (e) ESE(VGG-16 + FC4 × 2).

These results were obtained by computing the angular error (Equation (9)) between each pixel in the estimated per-pixel map and the groundtruth mask. It can be seen that all the jointly trained models significantly outperform the other

TABLE 1. Dice coefficient (Equation (8)) results of the models for the illuminant segmentation task. The names in parentheses show the base models used (in the case of our framework, the segmentation model is named first, followed by the estimation model). The best results are shown in bold. (Higher is better.)

Model	Mean	Median	Trimean	Best 25%	Worst 25%
Treshold [30]	0.73	0.73	0.73	0.86	0.58
Otsu [40]	0.80	0.83	0.82	0.95	0.63
Bianco-CNN [28]	0.79	0.78	0.78	0.93	0.66
Autoencoder-based [32] (VGG-16)	0.86	0.88	0.88	0.95	0.72
Autoencoder-based [32] (VGG-19)	0.82	0.84	0.83	0.93	0.66
T-large [31]	0.88	0.90	0.89	0.96	0.76
Seg [33]	0.89	0.91	0.90	0.97	0.76
VGG-16 [39]	0.88	0.90	0.90	0.96	0.76
ESE(Seg [33] + FC4)	0.89	0.92	0.92	0.97	0.77
ESE(Seg [33] + FC4x2)	0.89	0.92	0.92	0.97	0.77
ESE(VGG-16 + FC4)	**0.91**	**0.93**	**0.93**	**0.98**	**0.81**
ESE(VGG-16 + FC4x2)	0.90	0.93	0.92	0.97	0.79
RESE(VGG-16 + FC4x2)	**0.91**	**0.93**	0.92	0.97	0.80

TABLE 2. Angular error (Equation (9)) of the results of the models for the multi-illuminant estimation task. The names in parentheses show the base models used (in the case of our framework, the segmentation model is named first, followed by the estimation model, "x2" indicates two estimation models). The best results are shown in bold. (Lower is better.)

Model	Mean	Median	Trimean	Best 25%	Worst 25%
Gray-world [6]	5.55	5.37	5.45	2.54	8.84
1st-order Gray-Edge [8]	5.44	5.63	5.51	1.98	8.93
2nd-order Gray-Edge [8]	5.77	5.71	5.75	2.22	9.45
Sub-blocks Max-RGB [21]	5.71	4.94	5.12	2.57	10.17
ASM(Single) [26]	2.88	2.67	2.66	1.34	4.89
FC4(Squeezenet) [15]	4.09	3.85	3.91	1.78	6.76
ASM(Multi) [26]	22.72	23.14	22.72	10.18	35.40
MIRF [27]	7.08	6.20	6.40	2.70	12.67
Mixed-Illuminant CC [22]	8.25	7.59	7.74	3.81	13.67
Patch-based (White-Patch) [24]	4.3	2.89	-	1.02	10.13
Keypoint-based (White-Patch) [24]	5.46	3.59	-	1.11	13.15
Superpixel-based(2nd Order) [24]	4.2	3.1	-	1.09	9.32
DS-Net (HypNet+SelNet) [29]	6.31	3.95	-	0.85	15.95
Bianco-CNN [28]	4.65	4.42	4.42	2.63	7.27
Autoencoder-based [32] (VGG-16)	3.59	2.99	3.08	1.81	6.52
Autoencoder-based [32] (VGG-19)	4.38	3.82	3.99	2.45	7.26
Seg [33] + FC4x2	2.97	2.53	2.60	1.41	5.30
VGG-16 + FC4x2	2.98	2.51	2.59	1.45	5.33
ESE(Seg [33] + FC4)	2.66	2.17	2.25	1.18	5.00
ESE(Seg [33] + FC4x2)	2.56	**2.11**	2.21	**1.13**	4.74
ESE(VGG-16 + FC4)	2.58	**2.11**	2.21	1.22	4.75
ESE(VGG-16 + FC4x2)	**2.55**	2.12	**2.18**	1.19	**4.70**
RESE(VGG-16 + FC4x2)	2.64	2.23	2.31	1.24	4.81

models, with the largest margin of almost 0.5 degrees (14.5% improvement compared to the second best performing multi-illuminant model). It is also important to note that the smallest framework (composed of the small segmentation model [33] and shared reduced FC4 model) still outperforms other models. Furthermore, the models with the independent estimation estimators outperform their counterparts with shared estimators, at the cost of more parameters. This shows that different tradeoffs regarding accuracy, memory usage and speed can be implemented. Figure 4 shows the qualitative comparison of the segmentation and estimation results on images with two sources of illumination from our dataset.

Finally, we tested the performance of our framework on two commonly used single-illuminant datasets, the Cube+ [4] and ColorChecker [5] and compared it to other state-of-the-art methods. We show these results in Tables 3 and 4. The results show that, while some specialized single-illuminant learning-based models outperform our framework, it achieves by far the best results out of all tested multi-illuminant

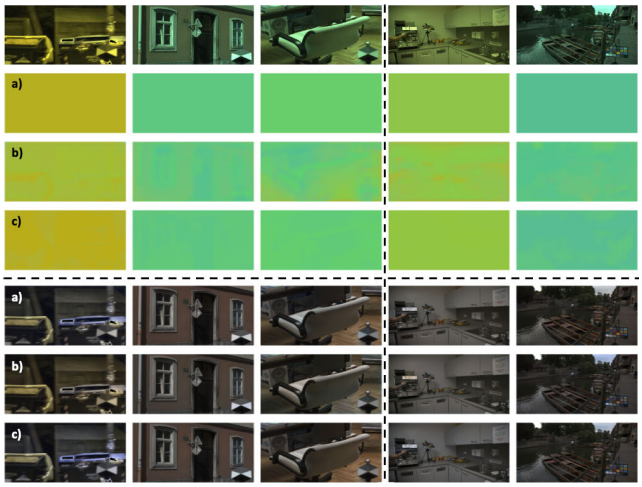

FIGURE 5. Example of the randomly selected images from the single-illuminant datasets, corresponding groundtruths, estimations and corrections. The first row is the input image, the first set of images are the per pixel illumination estimations, while the second set are the corrections. The first three columns correspond to images from the Cube+ [4] dataset, and the rest correspond to the ColorChecker [5] dataset. In each section, rows are marked with letters corresponding to different models and groundtruth. These are: (a) Groundtruth, (b) Autoencoder-based [32] (VGG-16), and (c) ESE(VGG-16 + FC4 × 2).

TABLE 3. Angular error (Equation (9)) of the results of the models for the single-illuminant estimation task on the Cube+ dataset [4]. The best results are shown in bold. The best performing multi-illuminant model is highlighted in yellow. Data for single-illuminant models was obtained from [18] (Lower is better).

Model	Mean	Median	Trimean	Worst 25%	Best 25%
White-Patch [7]	9.69	7.48	8.56	20.49	1.72
Gray-world [6]	7.71	4.29	4.98	20.19	1.01
Shades-of-gray [42]	2.59	1.73	1.93	6.19	0.46
1st-order Gray-Edge [8]	2.41	1.52	1.72	5.89	0.45
2nd-order Gray-Edge [8]	2.5	1.59	1.78	6.08	0.48
FFCC(model J) [20]	1.38	0.74	0.89	3.67	**0.19**
FC4(Squeezenet) [15]	1.35	0.93	1.01	3.24	0.3
Koscevic et. al.(VGG-16) [17]	1.34	0.83	0.97	3.2	0.28
MDLCC [43]	1.24	0.83	0.92	2.91	0.26
One-Net [18]	**1.21**	**0.72**	**0.83**	**3.05**	0.21
Autoencoder-based [32] (VGG-16)	4.21	3.99	4.03	7.18	1.85
Autoencoder-based [32] (VGG-19)	3.62	2.79	3.08	6.94	1.6
Bianco-CNN [28]	4.82	4.28	4.42	8.05	2.54
ESE(VGG-16 + FC4)	1.68	1.3	1.38	3.59	0.44
ESE(VGG-16 + FC4x2)	2.01	1.81	1.84	3.67	0.72

models. Furthermore, those results are still comparable with the best single-illuminant models, and the difference even in worst cases is less than the perceptual sensitivity of the human eye described in [41]. It is also important to note that all the single-illuminant models were trained on these datasets. However, our framework was trained on our multi-illuminant dataset and then only tested on these two single-illuminant datasets. This shows that our framework generalizes well over different images, as it is the only one of the multi-illuminant models that was able to achieve comparable results with the best single-illuminant models. Figure 5 provides a qualitative evaluation of the performance of our framework on single-illuminant datasets. It can be seen there that, even though the datasets are supposedly single-illuminant, some scenes do contain multiple illuminants, and that our model is capable of detecting this (second and last column).

TABLE 4. Angular error (Equation (9)) of the results of the models for the single-illuminant estimation task on the ColorChecker dataset [5]. The best results are shown in bold. The best performing multi-illuminant model is highlighted in yellow. Data for single-illuminant models was obtained from [15] and [16]. (Lower is better.)

Model	Mean	Median	Trimean	Worst 25%	Best 25%
White-Patch [7]	7.55	5.68	6.35	16.12	1.45
Gray-World [6]	6.36	6.28	6.28	10.58	2.33
1st-order Gray-Edge [8]	5.33	4.52	4.73	10.03	1.86
2nd-order Gray-Edge [8]	5.13	4.44	4.62	9.26	2.11
Shades-of-Gray [42]	4.93	4.01	4.23	10.2	1.14
Bayesian [12]	4.82	3.46	3.88	10.49	1.26
Exemplar based [10]	3.1	2.3	-	-	-
Regression Tree [13]	2.42	1.65	1.75	5.87	0.38
CRNA [16]	1.99	**1.01**	1.33	**3.2**	**0.22**
CCC (dist+ext) [19]	1.95	1.22	1.38	4.76	0.35
DS-Net (HypNet+SelNet) [29]	1.9	1.12	1.33	4.84	0.31
AlexNet-FC4 [15]	1.77	1.11	1.29	4.29	0.34
SqueezeNet-FC4 [15]	**1.65**	1.18	**1.27**	3.78	0.38
Autoencoder-based [32] (VGG-16)	4.78	4.29	4.41	7.92	2.41
Autoencoder-based [32] (VGG-19)	5.37	5.04	5.09	8.81	2.67
Bianco-CNN [28]	5.85	5.5	5.58	8.92	3.34
ESE(VGG-16 + FC4)	3.49	2.9	3.05	6.49	1.51
ESE(VGG-16 + FC4x2)	2.34	1.84	1.94	4.94	0.64

V. CONCLUSION

In this work, we presented a novel framework that is capable of segmenting and estimating illumination in scenes with one or two primary sources of illumination. The proposed framework is composed of specialized models for each task. First, a global estimation model is used to estimate the dominant illuminant in the scene. Then, a segmentation model is used to localize the influence of the estimated global illuminant. This produces regions of influence of illuminants, and the input image is masked using this segmentation. The masked images are then passed to estimation models that produce the estimation for those unmasked regions of the scene. The final estimation is done by linear combination of the estimated illuminants using the segmentation mask. Moreover, the proposed framework is modular as the estimation and segmentation models can easily be replaced, offering different tradeoffs in speed, memory, and accuracy.

The framework was tested on the novel dataset with 2500 images of varied indoor and outdoor scenes taken by 5 different cameras [3]. Our framework achieved the best results by a large margin, especially in the illuminant estimation task, with a 14.5% improvement above the second best scoring multi-illuminant model. We have also tested our framework on images from the Cube+ [4] and ColorChecker [5] single-illuminant datasets. For this task, we did not retrain the framework, but have used the best performing models from the multi-illuminant task. Here, our framework achieves excellent results, only slightly worse than specialized state-of-the-art single-illuminant estimation models. This shows the excellent generalization properties of our framework on cross dataset tasks.

REFERENCES

[1] M. Afifi and M. Brown, "What else can fool deep learning? Addressing color constancy errors on deep neural network performance," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 243–252.

[2] J. Von Kries, "Influence of adaptation on the effects produced by luminous stimuli," *Handbuch Der Physiologie Des Menschen*, vol. 3, pp. 109–282, Jan. 1905.

[3] I. Domislović, D. Vršnak, M. Subašić, and S. Loncarić. Accessed: Dec. 7, 2022. *Cube2: Large Mutli-Illuminant Dataset*. [Online]. Available: https://github.com/donikv/Cube2

[4] N. Banic and S. Loncaric, "Unsupervised learning for color constancy," 2017, *arXiv:1712.00436*.

[5] G. Hemrit, G. D. Finlayson, A. Gijsenij, P. Gehler, S. Bianco, B. Funt, M. Drew, and L. Shi, "Rehabilitating the colorchecker dataset for illuminant estimation," in *Proc. Color Imaging Conf.*, 2018, pp. 350–353.

[6] G. Buchsbaum, "A spatial processor model for object colour perception," *J. Franklin Inst.*, vol. 310, pp. 1–26, Jan. 1980.

[7] E. H. Land, "The retinex theory of color vision," *Sci. Amer.*, vol. 237, no. 6, pp. 108–129, 1977.

[8] J. Van De Weijer, T. Gevers, and A. Gijsenij, "Edge-based color constancy," *IEEE Trans. Image Process.*, vol. 16, no. 9, pp. 2207–2214, Sep. 2010.

[9] A. Gijsenij, T. Gevers, and J. Van De Weijer, "Generalized gamut mapping using image derivative structures for color constancy," *Int. J. Comput. Vis.*, vol. 86, no. 2, pp. 127–139, Jan. 2010.

[10] H. R. V. Joze and M. S. Drew, "Exemplar-based color constancy and multiple illumination," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 36, no. 5, pp. 860–873, May 2014.

[11] B. Funt and W. Xiong, "Estimating illumination chromaticity via support vector regression," in *Proc. Color Image Conf.*, vol. 50, Jan. 2004, pp. 47–52.

[12] P. V. Gehler, C. Rother, A. Blake, T. Minka, and T. Sharp, "Bayesian color constancy revisited," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[13] V. Agarwal, A. V. Gribok, A. Koschan, and M. A. Abidi, "Estimating illumination chromaticity via kernel regression," in *Proc. Int. Conf. Image Process.*, Oct. 2006, pp. 981–984.

[14] S. Bianco, C. Cusano, and R. Schettini, "Single and multiple illuminant estimation using convolutional neural networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4347–4362, Sep. 2017.

[15] Y. Hu, B. Wang, and S. Lin, "FC4: Fully convolutional color constancy with confidence-weighted pooling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jul. 2017, pp. 4085–4094.

[16] H.-H. Choi and B.-J. Yun, "Very deep learning-based illumination estimation approach with cascading residual network architecture (CRNA)," *IEEE Access*, vol. 9, pp. 133552–133560, 2021.

[17] K. Koscevic, M. Subasic, and S. Loncaric, "Iterative convolutional neural network-based illumination estimation," *IEEE Access*, vol. 9, pp. 26755–26765, 2021.

[18] I. Domislovic, D. Vrsnak, M. Subasic, and S. Loncaric, "One-Net: Convolutional color constancy simplified," *Pattern Recognit. Lett.*, vol. 159, pp. 31–37, Jul. 2022.

[19] J. T. Barron, "Convolutional color constancy," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 379–387.

[20] J. T. Barron and Y.-T. Tsai, "Fast Fourier color constancy," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 886–894.

[21] M. A. Hussain and A. S. Akbari, "Max-RGB based colour constancy using the sub-blocks of the image," in *Proc. 9th Int. Conf. Develop. eSyst. Eng. (DeSE)*, Aug. 2016, pp. 289–294.

[22] M. A. Hussain and A. S. Akbari, "Color constancy algorithm for mixed-illuminant scene images," *IEEE Access*, vol. 6, pp. 8964–8976, 2018.

[23] M. A. Hussain, A. Sheikh-Akbari, and E. A. Halpin, "Color constancy for uniform and non-uniform illuminant using image texture," *IEEE Access*, vol. 7, pp. 72964–72978, 2019.

[24] A. Gijsenij, R. Lu, and T. Gevers, "Color constancy for multiple light sources," *IEEE Trans. Image Process.*, vol. 21, no. 2, pp. 697–707, Feb. 2012.

[25] T. Akazawa, Y. Kinoshita, S. Shiota, and H. Kiya, "N-white balancing: White balancing for multiple illuminants including non-uniform illumination," *IEEE Access*, vol. 10, pp. 89051–89062, 2022.

[26] A. Akbarinia and C. A. Parraga, "Colour constancy beyond the classical receptive field," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 9, pp. 2081–2094, Sep. 2018.

[27] S. Beigpour, C. Riess, J. Van De Weijer, and E. Angelopoulou, "Multi-illuminant estimation with conditional random fields," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 83–96, Jan. 2014.

[28] S. Bianco, C. Cusano, and R. Schettini, "Color constancy using CNNs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 81–89.

[29] W. Shi, C. C. Loy, and X. Tang, "Deep specialized network for illuminant estimation," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 371–387.
[30] S.-H. Lee, S.-M. Woo, J.-H. Choi, and J.-O. Kim, "Two-step multi-illuminant color constancy for outdoor scenes," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 710–714.
[31] D. Vršnak, I. Domislović, M. Subašić, and S. Loncarić, "Illuminant estimation error detection for outdoor scenes using transformers," in *Proc. 12th Int. Symp. Image Signal Process. Anal. (ISPA)*, Sep. 2021, pp. 276–281.
[32] D. Vršnak, I. Domislović, M. Subašić, and S. Loncarić, "Autoencoder-based training for multi-illuminant color constancy," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 39, pp. 1076–1084, Jun. 2022.
[33] D. Vršnak, I. Domislović, M. Subašić, and S. Loncarić, "Illuminant segmentation for multi-illuminant scenes using latent illumination encoding," *Signal Process., Image Commun.*, vol. 108, Oct. 2022, Art. no. 116822.
[34] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2017, *arXiv:1608.03983*.
[35] S.-I. Amari, "Backpropagation and stochastic gradient descent method," *Neurocomputing*, vol. 5, nos. 4–5, pp. 185–196, 1993.
[36] Z. Li and Z. Ma, "Robust white balance estimation using joint attention and angular loss optimization," in *Proc. 13th Int. Conf. Mach. Vis.*, Jan. 2021, pp. 401–406.
[37] K. Zou, S. Warfield, A. Bharatha, C. Tempany, M. Kaus, S. Haker, W. Wells, F. Jolesz, and R. Kikinis, "Statistical validation of image segmentation quality based on a spatial overlap index," *Academic Radiol.*, vol. 11, pp. 89–178, Feb. 2004.
[38] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent. (MICCAI)* (Lecture Notes in Computer Science), vol. 9351. Cham, Switzerland: Springer, Nov. 2015, pp. 234–241.
[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2015, *arXiv:1409.1556*.
[40] C. Yu, C. Dian-ren, L. Yang, and C. Lei, "Otsu's thresholding method based on gray level-gradient two-dimensional histogram," in *Proc. 2nd Int. Asia Conf. Informat. Control, Autom. Robot. (CAR)*, Mar. 2010, p. 282.
[41] A. Gijsenij, T. Gevers, and M. P. Lucassen, "Perceptual analysis of distance measures for color constancy algorithms," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 26, no. 10, pp. 2243–2256, 2009.
[42] G. D. Finlayson and E. Trezzi, "Shades of gray and colour constancy," in *Proc. IST/SID Color Imag. Conf.*, vol. 1, Jan. 2004, pp. 37–41.
[43] J. Xiao, S. Gu, and L. Zhang, "Multi-domain learning for accurate and few-shot color constancy," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3258–3267.

DONIK VRŠNAK (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science, in 2018 and 2020, respectively. He is currently pursuing the Ph.D. degree in computing with the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. He is also doing research on color constancy, focused primarily on multi-illuminant segmentation and the perceptional properties of the human visual system related to color constancy. His research interests include image processing and analysis, bioinformatics, astronomy, and machine learning.

ILIJA DOMISLOVIĆ (Graduate Student Member, IEEE) received the B.Sc. and M.Sc. degrees in computer science, in 2018 and 2020, respectively. He is currently pursuing the Ph.D. degree in scientific field of computing (technical sciences) with the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. His research interests include image processing, image analysis, neural networks, and color constancy, with a focus on illumination estimation.

MARKO SUBAŠIĆ (Member, IEEE) received the Ph.D. degree from the Faculty of Electrical Engineering and Computing, University of Zagreb, in 2007. Since 1999, he has been working with the Department for Electronic Systems and Information Processing, Faculty of Electrical Engineering and Computing, University of Zagreb, where he is currently an Associate Professor. He teaches several courses at the graduate and undergraduate levels. His research interests include image processing and analysis and neural networks, with a particular interest in image segmentation, detection techniques, and deep learning. He is a member of the IEEE Computer Society, the Croatian Center for Computer Vision, the Croatian Society for Biomedical Engineering and Medical Physics, and the Centre of Research Excellence for Data Science and Advanced Cooperative Systems.

SVEN LONČARIĆ (Senior Member, IEEE) received the Ph.D. degree in electrical engineering from the University of Cincinnati, Cincinnati, OH, USA, in 1994. He was a Fulbright Scholar at the University of Cincinnati. He was an Assistant Professor with the New Jersey Institute of Technology, Newark, NJ, USA, from 2001 to 2003. He is currently a Professor of electrical engineering and computer science with the Faculty of Electrical Engineering and Computing, University of Zagreb, Croatia. He is also the Director of the Center for Computer Vision, University of Zagreb, the Head of the Image Processing Group, and the Co-Director of the Center of Excellence in Data Science and Cooperative Systems. He was the principal investigator on a number of research and development projects. He has coauthored more than 250 publications in scientific journals and conferences. His research interests include image processing and computer vision. He is a member of the Croatian Academy of Technical Sciences. He was the Chair of the IEEE Croatia Section. He received several awards for his scientific and professional work.