Deep Specialized Network for Illuminant Estimation

Wu Shi[1], Chen Change Loy[1,2], Xiaoou Tang[1,2]

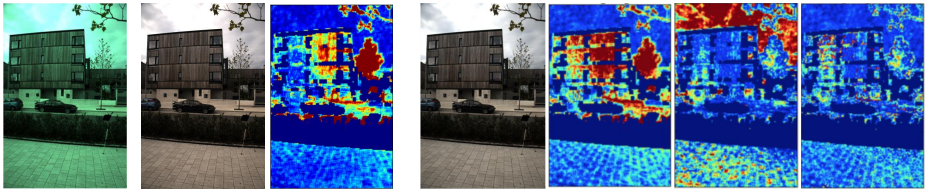
[1] Department of Information Engineering, The Chinese University of Hong Kong, Hong Kong, China
[2] Shenzhen Institutes of Advanced Technology, CAS, Shenzhen, China
{sw015, ccloy, xtang}@ie.cuhk.edu.hk

Abstract. Illuminant estimation to achieve color constancy is an ill-posed problem. Searching the large hypothesis space for an accurate illuminant estimation is hard due to the ambiguities of unknown reflections and local patch appearances. In this work, we propose a novel Deep Specialized Network (DS-Net) that is adaptive to diverse local regions for estimating robust local illuminants. This is achieved through a new convolutional network architecture with two interacting sub-networks, *i.e.* an *hypotheses network* (HypNet) and a *selection network* (SelNet). In particular, HypNet generates multiple illuminant hypotheses that inherently capture different modes of illuminants with its unique two-branch structure. SelNet then adaptively picks for confident estimations from these plausible hypotheses. Extensive experiments on the two largest color constancy benchmark datasets show that the proposed 'hypothesis selection' approach is effective to overcome erroneous estimation. Through the synergy of HypNet and SelNet, our approach outperforms state-of-the-art methods such as [1–3].

1 Introduction

The aim of color constancy is to recover the surface color under canonical (usually white) illumination from the observed color. Common computational approaches require estimating the spectral illumination of a scene to correct the extrinsic bias it induces. Illumination estimation can be understood as a process of searching through a hypothesis space to identify the best illuminant. It is often difficult to find a good one since the problem is underdetermined – both the illuminant and surface colors in an observed image are unknown. Finding a good hypothesis of illuminant becomes harder when there are ambiguities caused by complex interactions of extrinsic factors such as surface reflections and different texture appearances of objects.

Recent methods [2, 4] and [5] attempt to exploit the exceptional modelling capacity of convolutional network for this problem. We argue that it is still non-trivial to learn a model that can encompass the large and diverse hypothesis space given limited samples provided during the training stage. We believe that a model with higher flexibility could better handle ambiguous cases. The



(a) Input image (b) The restored image using CNN [2] and the corresponding pixel-wise angular error. (c) The restored image using DS-Net and the pixel-wise angular error maps for branch-1 and -2 for HypNet. The last image is the angular error map after the selection of SelNet.

Fig. 1. The proposed DS-Net shows superior performance over existing methods in handling regions with different intrinsic properties, thanks to the unique synergy between the hypotheses network (HypNet) and selection network (SelNet). In this example, the different branches of HypNet provide complementary illuminant estimations based on their specialization. SelNet automatically picks for the optimal estimations and yields a considerably lower angular error compared to that obtained from each respective branch in HypNet, as well as that obtained from CNN [2]. The angular error is the error in the iluminant estimate.

key principle advocated in [2, 4] and [5] is to arrange multiple layers of neurons to extract increasingly abstract features for reconstructing a restored image. We start from a similar principle but introduce new considerations in our network design for addressing the problem of illuminant estimation. The proposed network, named as *Deep Specialized Network* (DS-Net) consists of two closely coupled sub-networks.

1) *Hypotheses Network (HypNet)* – The sub-network learns to map an image patch to multiple hypotheses of illuminant of that patch. This is in contrast to existing network designs that usually provide just a single prediction. In our design, HypNet generates two competing hypotheses for an illuminant estimation of a patch through two branches that fork from a main CNN body. Each branch of HypNet is trained using a 'winner-take-all' learning strategy to automatically specialize to handle regions of certain appearance. For instance, as can be seen from Fig. 1, the first branch produces more accurate illuminant estimations for non-shadowed and bright regions like (*e.g.* sky), whilst the second branch is more effective on shadowed and textured areas (*e.g.* building and trees).

2) *Selection Network (SelNet)* – This sub-network makes an unweighted vote on the hypotheses produced by HypNet. Specifically, it takes an image patch and generates a score vector to pick the final illuminant hypothesis generated from one of the branches in HypNet. In other words, the SelNet acts like a 'filter', whose job is to decide which particular illuminant is more likely given the local patch statistics. We show that SelNet yields much robust final predictions than simply averaging the hypotheses. The entire structure of the two networks is shown in Fig. 2.

The main contribution of this study is a new deep specialized network effective for illuminant estimation. Specifically, we design a single network (*i.e.* a HypNet) to output multiple hypotheses, which resembles multiple expert net-

works in an ensemble. A diversity-encouraging 'winner-take-all' learning scheme is proposed to train the specialized network. We further present a viable way to design a separate network (*i.e.* a SelNet) for hypothesis selection. Extensive experiments on standard benchmarks show the superiority of DS-Net over existing methods in both global-illuminant of multi-illuminants estimation.

2 Related Work

Color constancy is a well-studied topic in both vision science and computer vision. There is a rich body of literature on illuminant estimation. These methods can be broadly divided into two categories: (1) statistic-based methods that estimate the illuminant based on image statistics or physical properties. These methods consider the relationship between color statistics and achromatic colors [6, 7], statistics inspired from the human visual system [8–12], spatial derivatives and frequency information from the image and scene illuminations [13–15], and specularity and shadows [16–18]; (2) learning-based approaches that estimate the illuminant using a model that is learned from training images. We refer readers to [19, 20] for excellent surveys. In this section, we highlight learning-based approaches that are related to our work.

In general, learning-based methods are shown to be more accurate than statistics-based approaches. Features considered in these studies are mostly handcrafted, including chromaticity histograms [21–24], full three-dimensional RGB histogram [25, 9], derivative and frequency features. Recent approaches have shown that relatively simple features, such as color and edge moments [26], or statistics of color chromaticity [1], could provide excellent performance.

While deep representation learned with CNN has achieved remarkable success in various high-level tasks [27–29] and a few low-level vision problems [30–35], it remains unclear if deep CNN can perform as well on the color constancy task. Barron [3] shows that his method can learn convolutional filters for accurate illuminant estimation. But he does not delve deeper into the use of deep CNNs. It is worth pointing out that Barron assumes that illuminant induces a global 2D translation in log-chrominance space. Such an assumption of uniform spectral distribution of light in an image may not work well in some common cases with multiple illuminants or scenes with in-shadow plus non-shadow regions. In contrast to [3], our approach does not assume single illuminant. We will show the effectiveness of the proposed approach over [3] on handling multi-illuminants.

Bianco *et al.* [2] make the first attempt to adopt a standard convolutional network for illuminant estimation. We show in the experiments that our network could provide more accurate estimates, thanks to the new network design with network-induced hypothesis selection. Under the global illuminant setting, while their method needs to specifically learn a separate support vector regressor to map local estimates to a global estimate, our approach can produce better results by just performing a simple median pooling on the already well-estimated illuminants from DS-Net.

A notable approach is proposed by Joze and Drew [36]. They adopt an exemplar-based approach - it finds similar surfaces in the training dataset, and estimating the illumination for each target surface through comparing the statistics of pixels belonging to similar surfaces with the target surface. This study shows the importance of capturing multiple modes through a non-parametric model. Our work is inspired by [36] but the proposed network does not require explicit nearest surface comparison. The multiple modes are inherently captured in the branch-level ensemble of HypNet.

There are other interesting approaches that exploit automatically detected objects such as faces [37] to guide the illuminant estimation. Approaches in [38, 39] require user guidance to deal with multiple illuminants.

3 Illuminant Estimation by Convolutional Network

Consider an image $\mathbf{I}_{\text{rgb}} = \{I_r, I_g, I_b\}$ taken from a linear RGB color camera with black level corrected and saturated pixels removed. The value of I_c for a Lambertian surface at pixel x is equal to the integral of the product of the illuminant spectral power distribution $E(x, \lambda)$, the surface reflectance $R(x, \lambda)$ and the sensor response function $S_c(\lambda)$:

$$I_c(x) = \int_{\Omega} E(x, \lambda) R(x, \lambda) S_c(\lambda) d\lambda, \qquad c \in \{r, g, b\}, \tag{1}$$

where λ is the wavelength, and Ω is the visible spectrum. From the Von Kries coefficient law [40], a simplified diagonal model is given by

$$I_c = E_c \times R_c, \qquad c \in \{r, g, b\}, \tag{2}$$

where E is the RGB illumination and R is the RGB value of reflectance under canonical (often white) illumination. Following this widely accepted model, the goal of color constancy is to estimate E from I, and then compute $R_c = I_c / E_c$.

Following existing studies [3, 41], we process images in the space of UV chrominance[3]. Specifically, we first convert the RGB channels of I to the log-homogeneous chrominance (I_u, I_v) defined as follows:

$$I_u = \log(I_r / I_g) \qquad I_v = \log(I_b / I_g), \tag{3}$$

and estimate the illumination in that space:

$$E_u = \log(E_r / E_g) \qquad E_v = \log(E_b / E_g). \tag{4}$$

One can easily recover (up to a scalar) the illumination E from UV to RGB [3] by following

$$E_r = \frac{\exp(-E_u)}{z} \quad E_g = \frac{1}{z} \quad E_b = \frac{\exp(-E_v)}{z} \tag{5}$$
$$z = \sqrt{\exp(-E_u)^2 + \exp(-E_v)^2 + 1}.$$

[3] As suggested by [41] and [3], the log-chrominance formulation is advantageous over the RGB formulation in that we have 2 unknown instead of 3, and R and I are related by simple linear constraint instead of a multiplicative constraint

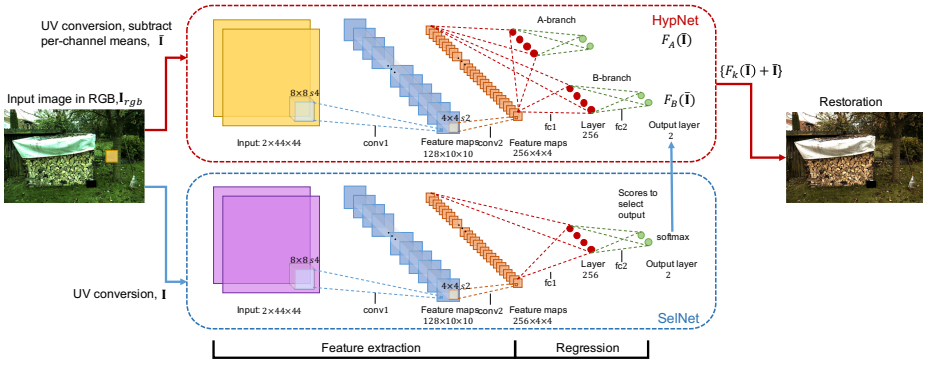


Fig. 2. The network architecture of Deep Specialized Network (DS-Net). It is trained to estimate illuminant of a given patch. It consists of two interacting networks, namely Hypotheses Network (HypNet) and Selection Network (SelNet). The former generates two hypotheses of illuminants from its two branches. The latter selects one of the hypotheses as the final estimation. It is possible to include more branches in HypNet.

In this study we present a deep convolutional network named as DS-Net for illuminant estimation. An illustration of the proposed DS-Net structure is given in Fig. 2. As introduced in Sec. 1, the proposed network is unique in that its two sub-networks, namely HypNet and SelNet, can interact to collectively provide accurate illuminant estimation. We will detail each sub-networks and their interaction as follows.

3.1 Hypothesis Network - A Branch-Level Ensemble Network

Hypothesis Network (HypNet) is trained, given a patch of image, to estimate multiple hypotheses of illuminant for that patch. The network consists of two stages:

1. **Feature extraction:** extracts spectral and spatial features from a UV patch of image.
2. **Regression:** estimates the illuminants from the features extracted by the previous stage.

We will show in the following that these two stages can be modelled by a convolutional neural network. The structure of the network is illustrated in Fig. 2.
Feature extraction. Previous color constancy methods considered both spectral and spatial information, such as the average of RGB, the color of edges, and the double-opponent response [9]. Barron [3] achieved state-of-art results by using extended spatial features. Chen *et al.* [42] applied discrete cosine transform (DCT) in log-space to extract illumination invariant features in face recognition. Following the literature, our model takes into account both spectral and spatial features. These features can be captured by convolving an image with a

bank of filters. The filters are learned during the training process of the network. Specifically, we use two convolutional layers and apply the Rectified Linear Unit ($\max(0, x)$) [43] on their outputs (see Fig. 2).

Regression. A straightforward method for regression is to use a stack of fully connected layers over the features from convolutional layers. However, we observe there are always some kinds of patches that the model cannot estimate well. We tried varying the number of layers, but the performance did not get any better. We conjecture that the difficulty may be due to the large complexity of the hypothesis space of this problem.

A plausible way to improve the performance of deep learning is to train an ensemble of neural networks and combine them during prediction. The benefits of ensemble methods are discussed in [44]. While the reasons combining models works so well are not fully comprehended, there is ample evidence that improvements over single models are the norm rather than the exception. The same observation has been frequently validated in many deep learning studies. For instance, Szegedy *et al.* [28] achieve top performance in ImageNet classification task through combining three residual and one inception network; DeepID2+ network [27] ensembles 25 networks for face verification. It is generally acknowledged that an ensemble is often much more accurate than the individual classifiers that make them up.

We wish to design a network that can covers a large and rich hypothesis space for improved performance. An ensemble network is a viable way to meet our objectives. To this end, we introduce a *branch-level ensemble* approach. Contrary to the convention of training multiple networks to form an ensemble, the proposed approach is implemented by forking after the last convolutional layer into two branches of fully connected layers, namely A-branch and B-branch[4]. Their mapping functions are represented as $F_{\mathrm{A}}(\cdot)$ and $F_{\mathrm{B}}(\cdot)$, respectively. The different branches constitute an ensemble. Such a design is computationally more attractive than a conventional network ensemble since they share common feature extraction layers. These two branches share only the input from the lower convolutional layer, but have individual parameters themselves and have no interactive connection. When the branches are trained with the 'winner-take-all' learning (discussed next), the two branches are able to cover different hypothesis spaces. As a result, the network will provide two intermediate hypotheses for any single patch.

To make a final decision for regression, two scores denoted as $\mathbf{s} = (s_{\mathrm{A}}, s_{\mathrm{B}})$, are given for the respective A and B branches and the branch with a higher score is selected to provide the output, *i.e.*, the scores serve as a filter to determine which signal could pass.

Data preprocessing. We subtract per-channel means of a patch from each channel, and finally add those means to the output. Specifically, the 2-channel input is denoted as $\bar{\mathbf{I}} = (I_u - \bar{I}_u, I_v - \bar{I}_v)$, where $(\bar{I}_u, \bar{I}_v)$ are per-channel means.

[4] We have tried more branches, but for this problem using more branches does not bring significant improvement. For efficiency and clarity, we present the two-branch version here.

The output are $F_{\mathrm{A}}(\bar{\mathbf{I}}) = (\tilde{E}_u - \bar{I}_u, \tilde{E}_v - \bar{I}_v)_{\mathrm{A}}$ for the A-branch and $F_{\mathrm{B}}(\bar{\mathbf{I}}) = (\tilde{E}_u - \bar{I}_u, \tilde{E}_v - \bar{I}_v)_{\mathrm{B}}$ for the B-branch, where $\tilde{\mathbf{E}} = (\tilde{E}_u, \tilde{E}_v) = F(\bar{\mathbf{I}}) + (\bar{I}_u, \bar{I}_v)$ is the final estimated illumination. This operation makes the performance of our model stable to a variety of illuminants. Please refer to the supplementary material for a detailed explanation.

***Winner-take-all* learning of HypNet.** In the training phase[5], a patch is extracted from an image and fed to the HypNet to obtain two hypotheses. The associated ground truth illuminant is provided. Then the score $(s_{\mathrm{A}}, s_{\mathrm{B}})$ for the branch whose hypothesis is closer to the ground truth is set to 1 and the other one to 0. We call these obtained scores as the *ground truth scores*, which will be used in SelNet training. Given a set of patches represented as $\{\bar{\mathbf{I}}\}$ and their corresponding ground truth illuminant $\{\mathbf{E}^*\}$, we use Euclidean loss[6] as the loss function to optimize HypNet. Specifically for each i-th patch, the loss is

$$L_i(\Theta) = \min_{k \in \{A,B\}} (||\tilde{\mathbf{E}}_i - \mathbf{E}_i^*||_2^2)_k, \tag{6}$$

where Θ represents the parameters of the convolutional layers and fully connected layers. The loss is minimized using stochastic gradient descent with the standard backpropagation. We adopt a batch-mode learning method with a batch size of 128.

Note that in our 'winner-take-all' learning scheme, only (the better) one of the branches is optimized and the other's forward signal and backward gradient are blocked[7]. In this way, at least one of the two branches is supposed to give a precise estimate and the two branches are able to complement each other to cover a larger hypothesis space. We attempted to back-propagate weighted sum errors to update the parameters of both branches but found that this scheme yielded much higher error in illuminant estimation.

In the test phase, the scores are obtained from another network, SelNet. We will introduce SelNet in the next section.

Discussion. We recommend using filters with a larger size in conv1 layer (see Fig. 2) to capture more spatial information. This follows several recent discoveries: (1) Barron [3] shows that using extended (spatial) features can improve their model by 10%-20%; (2) Gao *et al.* [9] demonstrates that using the structures analogous to the double-opponent cells in the human vision system will produce competitive results.

We note that there are different methods to create strong ensemble, *e.g.*, through enforcing interactions among the branches during training to increase diversity. We do not use deliberate method to create strong ensemble but just initialize the two branches differently with a similar spirit to random decision trees. Satisfactory performance is observed with this simple initialization approach, when it is used together with the proposed 'winner-take-all' learning.

[5] Implemented using Caffe [45].

[6] Despite the loss we use does not directly optimize the angular error typically employed in color constancy evaluation, satisfactory results are still observed.

[7] This scheme is also related to the Multiple Choice Learning [**?**].

3.2 Selection Network - A Hypothesis Selection Network

SelNet is trained to estimate the scores $\mathbf{s} = (s_A, s_B)$ that evaluate the quality of estimates, given the input patch and the hypotheses of that patch from HypNet. SelNet shares the same two-stage structure. However, the output of SelNet is not illuminant but a set of scores for the branches in HypNet. We apply a softmax operation on the output to get the scores normalized. Ideally, SelNet should give a higher score to the branch that is closer to the ground truth.

Input representation. We do not apply the data preprocessing of HypNet, since it may discard useful information such as local contrast. Consequently, we use the original patch in UV space. This representation only uses the information from the original data.

Learning for SelNet. In the training phase, an image patch and its ground truth illumination are extracted from an image. In addition, its two hypotheses and the ground truth scores are obtained from HypNet. We then arrange the input data in the corresponding form for SelNet and obtain an output from SelNet. The label is set to the ground truth scores. We optimize SelNet with multinomial logistic loss. In test phase, the output of SelNet is used to select one of the branches of HypNet.

3.3 Local to Global Estimation

Combining HypNet and SelNet, our DS-Net can predict patch-wise local illumination for an image. For the global-illuminant setting, a possible method is to learn a separate support vector regressor to aggregate local estimates to a global estimate [2]. Our approach can produce better results by simply performing a median pooling on all the local illuminant estimates of the image, without resorting to additional learning. Our unoptimized C++ code takes approximately 3 secs to process an image on a GPU.

4 Experiments

We evaluate the performance of our method in both global-illuminant and multi-illuminants settings in Sec. 4.1 and Sec. 4.2, respectively.

4.1 Global-Illuminant Setting

To evaluate the performance of our method in the global-illuminant setting, we use two standard datasets, *i.e.*, the Color Checker Dataset [25] reprocessed by Shi and Funt [46], and the NUS 8-camera dataset from Cheng *et al.* [47]. The Color Checker dataset contains 568 raw linear images with both indoor and outdoor scenes. The NUS 8-camera dataset from Cheng *et al.* consists of 1736 images from 8 different cameras, and about 210 individual scenes, where the same scene was photographed by each of the 8 cameras. For both of these datasets, the Macbeth Color Checker chart is placed in each image to estimate

Table 1. Performance comparison of the proposed DS-Net against various other methods on the Color Checker dataset [25, 46]. Some results were taken from past work therefore resulting in missing entries.

Methods	Mean	Median	Trimean	Best-25%	Worst-25%	95th percentile
White-Patch [40]	7.55	5.68	6.35	1.45	16.12	–
Edge-based Gamut [48]	6.52	5.04	5.43	1.90	13.58	–
Gray-World[6]	6.36	6.28	6.28	2.33	10.58	11.30
1st-order Gray-Edge [12]	5.33	4.52	4.73	1.86	10.03	11.00
2nd-order Gray-Edge [12]	5.13	4.44	4.62	2.11	9.26	–
Shades-of-Gray [49]	4.93	4.01	4.23	1.14	10.20	11.90
Bayesian [25]	4.82	3.46	3.88	1.26	10.49	–
General Gray-World [50]	4.66	3.48	3.81	1.00	10.09	–
Intersection-based Gamut [48]	4.20	2.39	2.93	0.51	10.70	–
Pixel-based Gamut [48]	4.20	2.33	2.91	0.50	10.72	14.10
Natural Image Statistics [10]	4.19	3.13	3.45	1.00	9.22	11.70
Bright Pixels [51]	3.98	2.61	–	–	–	–
Spatio-spectral (GenPrior) [52]	3.59	2.96	3.10	0.95	7.61	–
Cheng et al. [47]	3.52	2.14	2.47	0.50	8.74	–
Corrected-Moment (19 Color) [26]	3.50	2.60	–	–	–	8.60
Exemplar-based [36]	3.10	2.30	–	–	–	–
Corrected-Moment (19 Edge) [26]	2.80	2.00	–	–	–	6.90
CNN [2]	2.36	1.98	–	–	–	–
Regression Tree [1]	2.42	1.65	1.75	0.38	5.87	–
CCC (disc+ext) [3]	1.95	1.22	1.38	0.35	**4.76**	**5.85**
HypNet One Branch	2.18	1.35	1.54	0.38	5.42	6.69
HypNet (A-branch)	5.06	4.38	4.52	1.26	10.05	12.43
HypNet (B-branch)	4.55	2.35	3.10	0.50	12.21	15.50
DS-Net (Average)	3.74	2.99	3.18	0.86	7.83	9.27
DS-Net (HypNet+SelNet)	**1.90**	**1.12**	**1.33**	**0.31**	4.84	5.99
DS-Net (HypNet+Oracle)	1.15	0.76	0.86	0.22	2.72	3.35

the ground truth illuminant. The color checker chart is masked out during the training and evaluation. Our model is learned and evaluated using a three-fold cross-validation. The angular error between the estimated illuminant $\tilde{E}_{rgb}$ and the ground truth illuminant E^*_{rgb} is computed for each image:

$$\epsilon = \arccos\left(\frac{\tilde{E}_{rgb} \cdot E^*_{rgb}}{\|\tilde{E}_{rgb}\| \cdot \|E^*_{rgb}\|}\right). \tag{7}$$

We report the following metrics following existing studies [3, 1]: the mean, the median, the tri-mean, the means of the lowest-error 25% and the highest-error 25% of the data, and the 95 percentile for the Color Checker dataset. For the NUS 8-camera dataset, we run 8 different experiments on the subset for each camera, and report the geometric mean of each error metric for all the methods. A number of different color constancy algorithms are compared, and the reported baseline results were taken from past papers [3]. Experimental results of the Color Checker dataset and NUS 8-camera dataset are summarized in Tables 1 and 2, respectively.

Comparison with state-of-the-arts. On both the Color Checker and NUS 8-camera datasets, the proposed method 'DS-Net (HypNet+SelNet)' achieves the lowest mean and median errors in comparison to existing methods, including the CNN method presented in [2]. We show some examples of our performance against competitive methods in Fig. 3. In comparison to existing approaches, it

Table 2. Performance comparison of the proposed DS-Net against various other methods on the Cheng et al. [47] dataset.

Methods	Mean	Median	Trimean	Best-25%	Worst-25%
White-Patch [40]	10.62	10.58	10.49	1.86	19.45
Edge-based Gamut [48]	8.43	7.05	7.37	2.41	16.08
Pixel-based Gamut [48]	7.70	6.71	6.90	2.51	14.05
Intersection-based Gamut [48]	7.20	5.96	6.28	2.20	13.61
Gray-World[6]	4.14	3.20	3.39	0.90	9.00
Bayesian [25]	3.67	2.73	2.91	0.82	8.21
Natural Image Statistics [10]	3.71	2.60	2.84	0.79	8.47
Shades-of-Gray [49]	3.40	2.57	2.73	0.77	7.41
Spatio-spectral (ML) [52]	3.11	2.49	2.60	0.82	6.59
General Gray-World [50]	3.21	2.38	2.53	0.71	7.10
2nd-order Gray-Edge [12]	3.20	2.26	2.44	0.75	7.27
Bright Pixels [51]	3.17	2.41	2.55	0.69	7.02
1st-order Gray-Edge [12]	3.20	2.22	2.43	0.72	7.36
Spatio-spectral (GenPrior) [52]	2.96	2.33	2.47	0.80	6.18
Cheng et al. [47]	2.92	2.04	2.24	0.62	6.61
CCC (disc+ext) [3]	2.38	1.48	1.69	**0.45**	5.85
Regression Tree [1]	2.36	1.59	1.74	0.49	**5.54**
HypNet One Branch	2.56	1.87	2.01	0.51	6.46
HypNet (A-branch)	3.49	2.94	3.03	0.90	7.00
HypNet (B-branch)	5.17	2.91	3.50	0.91	13.03
DS-Net (Average)	3.41	2.36	2.72	0.73	7.69
DS-Net (HypNet+SelNet)	**2.24**	**1.46**	**1.68**	0.48	6.08
DS-Net (HypNet+Oracle)	1.32	0.93	1.01	0.33	2.97

is observed that our method performs better on complex and diverse regions, *e.g.* texture areas such as grass field, or smooth regions such as wall or sky. The results suggest the effectiveness of adopting a branch-level ensemble network with hypothesis selection.

Ablation analysis. We evaluated different variants of DS-Net:

- **HypNet One Branch** - it is a variant without the branch-level ensemble, i.e., it is a normal network with only one branch of fully connected layers, so SelNet is not needed for hypothesis selection.
- **HypNet (A-branch) or (B-branch)** - these variants refer to a model that generate estimations based on either A-branch or B-branch of HypNet.
- **DS-Net (Average)** - this variant generates an estimation by averaging the hypotheses from both A-branch and B-branch. It represents the typical way of generating predictions from an ensemble.
- **DS-Net (HypNet+SelNet)** - this is our full model with hypothesis selection using SelNet.
- **DS-Net (HypNet+Oracle)** - in this variant the hypothesis is selected by an oracle. The oracle selects the branch of which the estimation is closest to the ground truth.

From the results on the Color Checker dataset, it is observed that the mean and median errors are reduced by 13% (from 2.18 to 1.90) and 17% (from 1.35 to 1.12), respectively by using the branch-level ensemble in comparison to the variant 'HypNet One Branch'. It is interesting to point out that none of the two branches individually can achieve satisfactory performance. Averaging the hypotheses, *i.e.* HypNet (Average), does not improve the performance either.

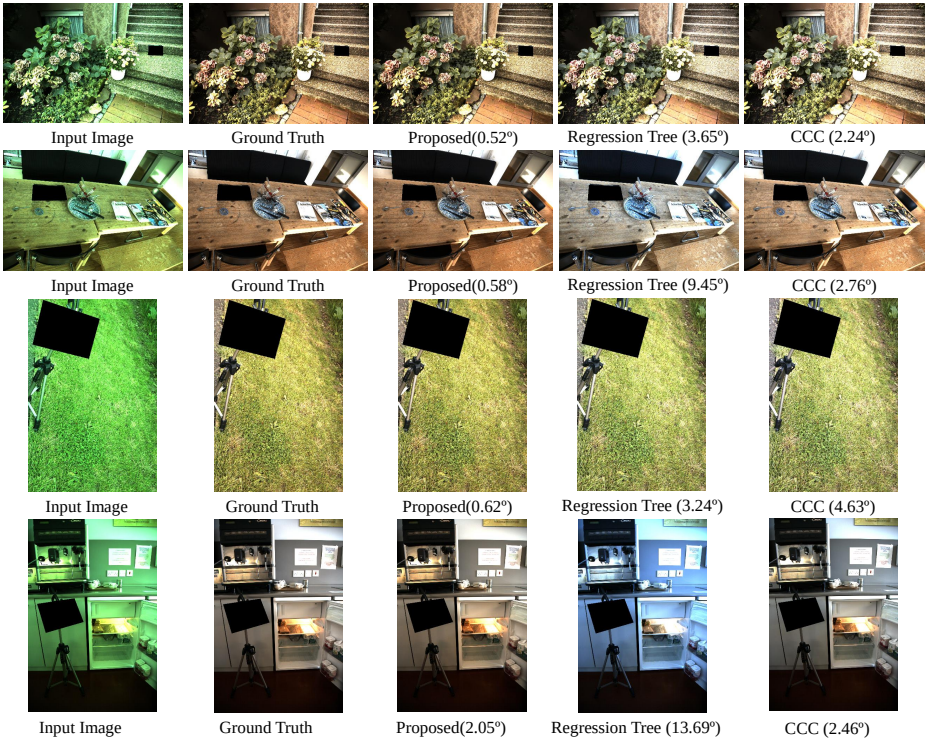


Fig. 3. *Global illuminant setting*: Restored images from the Color Checker dataset using the illuminants estimated from three different methods including the proposed DS-Net, Regression Tree [36] and CCC [3]. The angular error is provided at the bottom of each image. We follow [1] to apply gamma function on RAW images for better visualization.

However, a significant improvement is obtained when SelNet is used for hypothesis selection. Note that if the best branch is selected by an oracle, the errors can be further reduced by a large margin (39% and 32% of the mean and median errors). The results suggest the large potential of hypothesis selection and there is still a room for further optimization. Our current SelNet achieves a selection accuracy of 75%-77% on the test folds. In Fig. 4, we show two examples of angular error maps obtained by using different variants of DS-Net. Fig. 5 illustrates the evolution of illumination estimated by the two branches of HypNet. It is observed that both branches are gradually converging to the ground truth and, at the same time, preserving their own specialities.

We also perform evaluation on SelNet by testing it with different input representations, *i.e.* with and without per-channel means subtraction. It is observed that with per-channel means subtraction the selection accuracy of SelNet drops to 67%, leading to higher mean and median errors in the final illuminant estimation (3.81 and 2.80).

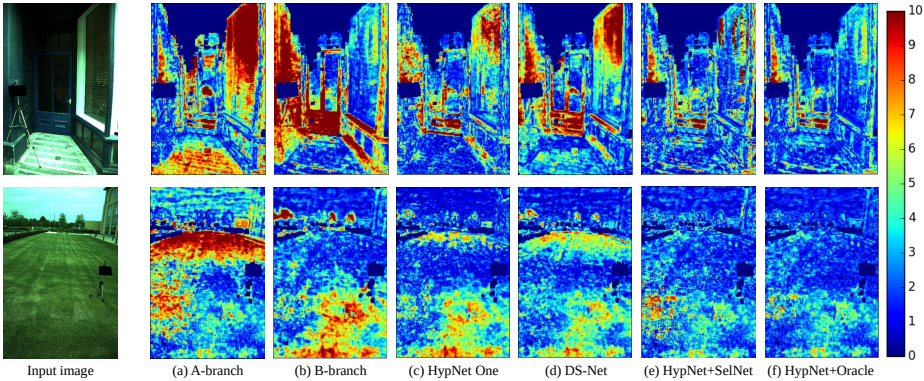


Fig. 4. *Global illuminant setting*: (a, b) The respective per-pixel angular error map of A-branch and B-branch of HypNet. (c) HypNet One Branch. (d) DS-Net (Average). (e) The full model DS-Net (HypNet+SelNet). (f) Upper bound DS-Net (HypNet+Oracle).

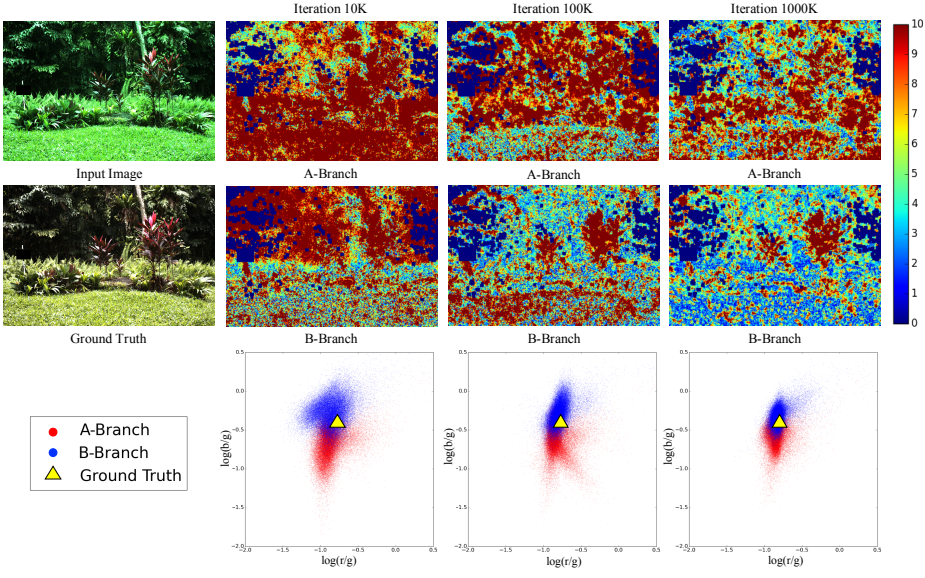


Fig. 5. *Global illuminant setting*: Column 1: the input image and the restored image using ground truth illumination. Column 2-4: the first two rows show the per-pixel angular error map of A-branch and B-branch of HypNet, using the models after 10K, 100K and 1000K training iterations. The last row depicts the UV chrominance of per-pixel illumination estimated by the two branches at the corresponding iterations.

4.2 Multi-Illuminant Setting

The proposed DS-Net by nature predicts patch-wise local illumination for an image. Thus it is capable of dealing with multi-illuminant settings although we do not introduce specific mechanisms, *e.g.* segmentation [53], to handle the different illuminants.

In this section, we evaluate the performance of our method on the popular outdoor multi-illuminant dataset proposed by Gijsenij *et al.* [53]. This dataset contains 9 challenging outdoor images with two illuminants for each image. Pixel-wise ground truth illuminants are provided for evaluation. The per-image error metric is the mean of pixel-wise angular error. Following [53], we report the mean and median errors of all images in the whole dataset. Considering the limited number of test images, global illuminant baselines and our method are first pre-trained on the Color Checker dataset, and then tested on the outdoor dataset. This also makes the task more challenging due to the cross-dataset evaluation.

We report the results of two state-of-art multi-illuminant methods, namely the Multiple Light Sources [53] (using White Patch and Gray World) and the Exemplar-Based [36] method using surface estimates, together with two state-of-art global-illuminant methods, Regression Tree [1] and CCC (dist+ext) [3]. The results of Multiple Light Sources and Exemplar-Based methods were obtained from the original paper [53] and [36]. We obtained the codes of Regression Tree [1] from its project page and retrained it on the Color Checker dataset. We reimplemented CCC (dist+ext) [3]. We ensure that both methods achieve comparable performance to their reported results under the global illuminant setting.

The results are summarized in Table 3. It is observed that the proposed DS-Net outperforms existing global methods by a significant margin. Our approach also reports competitive performance in comparison to state-of-art multi-illuminant methods [53, 36]. Note that unlike state-of-the-art exemplar-based method [36] that requires finding surfaces for both training and test images by mean-shift segmentation, and storing surfaces of all training images for nearest neighbor comparison, our approach only needs to perform pure feed-forward test given a new image. Qualitative results are shown in Fig. 6.

Table 3. Performance comparison of the proposed DS-Net against various other methods on the multi-illuminant outdoor dataset [53].

Methods	Mean	Median
Global state-of-the-arts:		
Regression Tree [1]	9.3	7.8
CCC (disc+ext) [3]	8.4	9.0
Multi-illuminant state-of-the-arts:		
Multiple Light Sources + White-Patch [53]	-	6.7
Multiple Light Sources + Gray-World [53]	-	6.4
Exemplar-Based Multi [36]	-	**4.3**
DS-Net (HypNet+SelNet)	4.8	4.6

Fig. 6. *Multi-illuminants setting*: Results from the outdoor multi-illuminant image dataset [53]. For each group, the images, from left to right, are respectively: the original image, the ground-truth pixel-wise illumination for that image, the estimated results using DS-Net (HypNet+SelNet), CCC [3], and Regression Tree [36]. Best viewed in color.

5 Conclusion

We have presented a new Deep Specialized Network (DS-Net) for illuminant estimation. The proposed network uniquely combines two networks: a multi-hypotheses network (HypNet) and a hypothesis selection network (SelNet), to work hand-in-hand for robust estimation. A novel notion of 'branch-level ensemble' is introduced. Through the proposed diversity-encouraging winner-take-all learning scheme, we observed that the two branches of HypNet automatically specialize on estimating illuminants for specific regions. When this capability is coupled with SelNet, state-of-the-art performances are achieved on the two largest color constancy dataset. Future work will investigate more effective selection scheme for a larger ensemble. In addition, it will be interesting to explore the applicability of specialized network for high-level vision task.

Acknowledgment. This work is partially supported by SenseTime Group Limited.

References

1. Cheng, D., Price, B., Cohen, S., Brown, M.S.: Effective learning-based illuminant estimation using simple features. In: IEEE Conference on Computer Vision and Pattern Recognition. (2015) 1000–1008

2. Bianco, S., Cusano, C., Schettini, R.: Single and multiple illuminant estimation using convolutional neural networks. arXiv preprint arXiv:1508.00998 (2015)
3. Barron, J.T.: Convolutional color constancy. In: IEEE International Conference on Computer Vision. (2015) 379–387
4. Bianco, S., Cusano, C., Schettini, R.: Color constancy using cnns. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. (2015) 81–89
5. Lou, Z., Gevers, T., Hu, N., Lucassen, M.: Color constancy by deep learning. In: British Machine Vision Conference. (2015)
6. Buchsbaum, G.: A spatial processor model for object colour perception. Journal of the Franklin institute **310**(1) (1980) 1–26
7. Land, E.H., McCann, J.J.: Lightness and retinex theory. Journal of the Optical Society of America A **61**(1) (1971) 1–11
8. Gao, S., Han, W., Yang, K., Li, C., Li, Y.: Efficient color constancy with local surface reflectance statistics. In: European Conference on Computer Vision. (2014) 158–173
9. Gao, S., Yang, K., Li, C., Li, Y.: A color constancy model with double-opponency mechanisms. In: IEEE International Conference on Computer Vision. (2013) 929–936
10. Gijsenij, A., Gevers, T.: Color constancy using natural image statistics and scene semantics. IEEE Transactions on Pattern Analysis and Machine Intelligence **33**(4) (2011) 687–698
11. Gijsenij, A., Gevers, T., Van De Weijer, J.: Improving color constancy by photometric edge weighting. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(5) (2012) 918–929
12. Van De Weijer, J., Gevers, T., Gijsenij, A.: Edge-based color constancy. IEEE Transactions on Image Processing **16**(9) (2007) 2207–2214
13. Bianco, S., Ciocca, G., Cusano, C., Schettini, R.: Improving color constancy using indoor–outdoor image classification. IEEE Transactions on Image Processing **17**(12) (2008) 2381–2392
14. Bianco, S., Ciocca, G., Cusano, C., Schettini, R.: Automatic color constancy algorithm selection and combination. Pattern Recognition **43**(3) (2010) 695–705
15. Drew, M.S., Funt, B.V.: Variational approach to interreflection in color images. Journal of the Optical Society of America A **9**(8) (1992) 1255–1265
16. Drew, M.S., Joze, H.R.V., Finlayson, G.D.: Specularity, the zeta-image, and information-theoretic illuminant estimation. In: European Conference on Computer Vision Workshop. (2012) 411–420
17. Lee, H.C.: Method for computing the scene-illuminant chromaticity from specular highlights. Journal of the Optical Society of America A **3**(10) (1986) 1694–1699
18. Tan, R.T., Nishino, K., Ikeuchi, K.: Color constancy through inverse-intensity chromaticity space. Journal of the Optical Society of America A **21**(3) (2004) 321–334
19. Hordley, S.D.: Scene illuminant estimation: past, present, and future. Color Research & Application **31**(4) (2006) 303–314
20. Gijsenij, A., Gevers, T., Van De Weijer, J.: Computational color constancy: Survey and experiments. IEEE Transactions on Image Processing **20**(9) (2011) 2475–2489
21. Cardei, V.C., Funt, B., Barnard, K.: Estimating the scene illumination chromaticity by using a neural network. Journal of the Optical Society of America A **19**(12) (2002) 2374–2386

22. Finlayson, G.D., Hordley, S.D., Hubel, P.M.: Color by correlation: A simple, unifying framework for color constancy. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(11) (2001) 1209–1221
23. Funt, B., Xiong, W.: Estimating illumination chromaticity via support vector regression. In: Color and Imaging Conference. Volume 2004. (2004) 47–52
24. Rosenberg, C., Hebert, M., Thrun, S.: Color constancy using KL-divergence. In: IEEE International Conference on Computer Vision. Volume 1. (2001) 239–246
25. Gehler, P.V., Rother, C., Blake, A., Minka, T., Sharp, T.: Bayesian color constancy revisited. In: IEEE Conference on Computer Vision and Pattern Recognition. (2008) 1–8
26. Finlayson, G.: Corrected-moment illuminant estimation. In: IEEE International Conference on Computer Vision. (2013) 1904–1911
27. Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust. In: IEEE Conference on Computer Vision and Pattern Recognition. (2015) 2892–2900
28. Szegedy, C., Ioffe, S., Vanhoucke, V.: Inception-v4, inception-resnet and the impact of residual connections on learning. arXiv preprint arXiv:1602.07261 (2016)
29. Zhu, S., Liu, S., Loy, C.C., Tang, X.: Deep cascaded bi-network for face hallucination. In: European Conference on Computer Vision. (2016)
30. Dong, C., Loy, C.C., He, K., Tang, X.: Image super-resolution using deep convolutional networks. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(2) (2015) 295–307
31. Cui, Z., Chang, H., Shan, S., Zhong, B., Chen, X.: Deep network cascade for image super-resolution. In: European Conference on Computer Vision. (2014) 49–64
32. Xu, L., Ren, J.S., Liu, C., Jia, J.: Deep convolutional neural network for image deconvolution. In: Advances in Neural Information Processing Systems. (2014) 1790–1798
33. Dong, C., Loy, C.C., Tang, X.: Accelerating the super-resolution convolutional neural network. In: European Conference on Computer Vision. (2016)
34. Hui, T.W., Loy, C.C., Tang, X.: Depth map super resolution by deep multi-scale guidance. In: European Conference on Computer Vision. (2016)
35. Xie, J., Xu, L., Chen, E.: Image denoising and inpainting with deep neural networks. In: Advances in Neural Information Processing Systems. (2012) 341–349
36. Joze, H.R.V., Drew, M.S.: Exemplar-based color constancy and multiple illumination. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(5) (2014) 860–873
37. Bianco, S., Schettini, R.: Adaptive color constancy using faces. IEEE Transactions on Pattern Analysis and Machine Intelligence **36**(8) (2014) 1505–1518
38. Hsu, E., Mertens, T., Paris, S., Avidan, S., Durand, F.: Light mixture estimation for spatially varying white balance. In: ACM Transactions on Graphics. Volume 27. (2008) 70
39. Boyadzhiev, I., Bala, K., Paris, S., Durand, F.: User-guided white balance for mixed lighting conditions. ACM Transactions on Graphics **31**(6) (2012) 200
40. Brainard, D.H., Wandell, B.A.: Analysis of the retinex theory of color vision. Journal of the Optical Society of America A **3**(10) (1986) 1651–1661
41. Finlayson, G.D., Drew, M.S., Lu, C.: Intrinsic images by entropy minimization. In: European conference on computer vision, Springer (2004) 582–595
42. Chen, W., Er, M.J., Wu, S.: Illumination compensation and normalization for robust face recognition using discrete cosine transform in logarithm domain. IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics **36**(2) (2006) 458–466

43. Nair, V., Hinton, G.E.: Rectified linear units improve restricted boltzmann machines. In: International Conference on Machine Learning. (2010) 807–814
44. Dietterich, T.G.: Ensemble methods in machine learning. In: Multiple classifier systems. Springer (2000) 1–15
45. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: ACM Multimedia. (2014) 675–678
46. Shi, L., Funt, B.: Re-processed version of the gehler color constancy dataset of 568 images. accessed from http://www.cs.sfu.ca/ colour/data/
47. Cheng, D., Prasad, D.K., Brown, M.S.: 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) (2014) 1049–1058
48. Barnard, K.: Improvements to gamut mapping colour constancy algorithms. In: European Conference on Computer Vision. (2000) 390–403
49. Finlayson, G.D., Trezzi, E.: Shades of gray and colour constancy. In: Color and Imaging Conference. Volume 2004. (2004) 37–41
50. Barnard, K., Martin, L., Coath, A., Fun, B.: A comparison of computational color constancy algorithms. ii. experiments with image data. IEEE Transactions on Image Processing **11**(9) (2002) 985–996
51. Joze, H.R.V., Drew, M.S., Finlayson, G.D., Rey, P.A.T.: The role of bright pixels in illumination estimation. In: Color and Imaging Conference. Volume 2012. (2012) 41–46
52. Chakrabarti, A., Hirakawa, K., Zickler, T.: Color constancy with spatio-spectral statistics. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(8) (2012) 1509–1519
53. Gijsenij, A., Lu, R., Gevers, T.: Color constancy for multiple light sources. IEEE Transactions on Image Processing **21**(2) (2012) 697–707