

Light Direction and Color Estimation from Single Image with Deep Regression

Hassan A. Sial¹, Ramon Baldrich¹, Maria Vanrell¹, Dimitris Samaras²

¹Computer Vision Center, Universitat Autnoma de Barcelona, Barcelona, Spain

²Department of Computer Science, Stony Brook University, Stony Brook, New York, USA

Abstract

We present a method to estimate the direction and color of a scene light source from a single image. Our method is based on two main ideas: (a) we use a new synthetic dataset with strong shadow effects with similar constraints to SID dataset; (b) we define a deep architecture trained on the mentioned dataset to estimate direction and color of the scene light source. Apart from showing a good performance on synthetic images, we additionally propose a preliminary procedure to obtain light positions of the Multi-Illumination dataset, and, in this way, we also prove that our trained model achieves a good performance when it is applied to real scenes.

Introduction

Scene appearance is directly dependent on the light source properties, such as the spectral composition of emitted light, and the position and direction of the light source, whose interaction with the scene objects provoke shaded surfaces or dark cast shadows that become essential visual cues to understand image content. Estimating the properties of the light conditions from a single image is an initial step to improve subsequent computer vision algorithms for image understanding. In this paper we perform a preliminary study to estimate color and position of light in a simple and unified approach, that is based on the shading properties of the image where we assume a single scene illuminant.

Estimating the color of the light from a single image has been focus of attention in previous research. Computational color constancy (CC) has been studied in a large number of works [8, 7, 12] where the problem was tackled from different points of views [12]. A first approach was to extract statistics from RGB image values under different assumptions to estimate the canonical white. A second approach was to introduce spatio-statistical information like gradients or frequency content of the image. One last group of CC algorithms was to try to get the information from physical cues of the image (highlights, shadows, inter-reflections, etc). In the last years, new approaches have been based on deep learning frameworks where the solution is driven by the data with physical constrains in the loss functions. An updated comprehensive compendium and comparison of CC algorithm performances can be found in [4, 16, 5, 27]. In this work we propose one more approach based on a deep architecture, but color of the light source is jointly estimated with the light direction.

Estimating the direction of the light has also been tackled from different areas like, computer graphics, computer vision or augmented reality. Single image light direction estimation can be divided in two different kinds of approaches. First, those in which light probes with known reflectance and geometric properties are used. A Specular sphere is commonly used to represent light position in different computer graphics applications [6, 1, 22]. But, random shaped objects to detect light position

were used by Mandl et-al in [17], jointly with a deep learning approach to get the light position with each of these random shaped object. Second, we find those works in which no probe is used, and where multiple image cues such as shading, shape and cast shadows are the basis to estimate light direction. Some examples of these works can be found in computer vision literature [13, 21, 20, 19, 2].

More recently, some deep learning methods have been proposed to estimate scene illumination and have been used for different computer graphics tasks. Gardner et al.[10] introduced a method to convert low dynamic range (LDR) indoor images to high dynamic range (HDR) images, first they used a deep network to localize the light source in LDR image environmental map and then they used another network with these annotated LDR images to convert them to HDR image. Following a similar approach, Geoffroy et al. [11] introduced a method to convert outdoor LDR images to HDR images. They trained their network with a set of panorama images and predicted HDR environmental maps with sky and sun positions. Later on, Geoffroy et al [9] extended their previous idea for indoor lighting but replaced the environmental maps with light geometric and photometric properties. Sun et al. [24] introduced an encoder-decoder based network to relight a single portrait image, the encoder predicts the input image environmental map and an embedding for the scene information, while the decoder builds a new version of the image scene with the target environmental map and obtains a relighted portrait. Very recently, Muramann et al. [18] have introduced the *Multi-illuminant dataset* of 1000 scenes each one acquired under 25 different light position conditions, and they used a deep network to predict a right sided illuminated image from its corresponding left sided illuminated image. In this work we will test our proposal on this new wild dataset after providing a procedure to compute the light direction from each sample.

To sum up, we can state that a large range of works have tackled the problem of estimating color and direction of the scene light source from different points of views and focusing on specific applications. In this work we propose an easy end-to-end approach to jointly characterize the light source of a scene, both for color and direction. We pursuit to measure the level of accuracy we can achieve, in order it can be applied to a wide range of images, without using probes in the scene and becoming a robust preliminary stage to be subsequently combined with any task. To this end, the paper is organized as follows: first we introduce a new synthetic dataset, secondly we use it to train a deep architecture that estimates light properties from a single image, finally we show how our proposal performs on three different scenarios, synthetic, real indoor and natural images.

A synthetic dataset

In order to train our end-to-end network that estimates light conditions we developed a new image dataset similar to SID

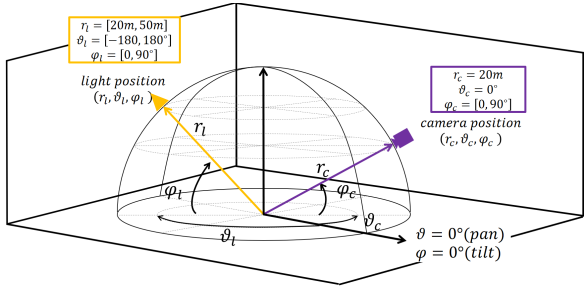


Figure 1. Image Generation Setup. Camera and light positions are given in spherical coordinates (r, ϑ, φ) .

[23], which was created for intrinsic image decomposition. This dataset is formed by a large set of images of surreal scenes where Shapenet objects [3] are enclosed in the center of multi-sided rooms with highly diverse backgrounds provided by flat textures on the room walls, resulting in a large range of different light effects. From now on we will refer to this dataset as SID1, and we will propose a new one better adapted to our problem, using the same methodology and software provided by the authors, which is based on an open source Blender rendering engine to synthesize images.

Our new dataset will be called SID2 dataset. The main difference is that it introduces more than one object in each scene, with the aim of increasing the number of strong light effects and interactions. Additionally, we also introduce more variability in the distance from the light source to the scene center. The dataset is formed by 45,000 synthetic images with the corresponding ground truth data: direction and color of the light source.

We did several assumptions in building the dataset: (a) objects are randomly positioned around the scene center but always close to the room ground floor to have realistic cast shadows; (b) light source direction goes from scene center to a point onto an imaginary semi-sphere of a random radius and with random RGB color; (c) camera is randomly positioned at a random distance from the center of the scene and always with the focal axis pointing to the scene center. In Figure 1 we show the diagram of the synthetic world we defined for the generation of SID2. More specifically, we took 45,000 3D objects from Shapenet dataset [3]. Likewise in [23] we did not use textured surfaces, we used a diffuse bidirectional scattering distribution function (BSDF) with random color and roughness values for each mesh texture in each object. This roughness parameter controls how much light is reflected back from each object surface. We randomly picked from 1 to 3 objects in each image. They were placed at random locations within the camera view range. We placed an empty object in the center of the scene to ensure non-overlapping between the rest of objects. Light direction was randomly defined in spherical coordinates $(r_l, \vartheta_l, \varphi_l)$, being radius, pan and tilt, respectively. We took random values within the ranges of $[20m, 50m]$ $[30^\circ, 90^\circ]$ and $[0^\circ, 360^\circ]$ respectively, in steps of 1° for pan and 5° for tilt. Light intensity and chromaticity was randomly selected, but chromaticity was constrained to be around the Planckian locus to simulate natural lighting conditions. Camera position is also denoted in spherical coordinates as $(r_c, \vartheta_c, \varphi_c)$, where r_c was fixed at $20m$ and pan, $\vartheta_c = 0^\circ$, the tilt range randomly varied within $[10^\circ, 70^\circ]$. In the final ground-truth (GT), light pan and tilt are provided with reference to the camera position, in order not to depend on real world positions which are usually not available in real images. Backgrounds were generated in the same way as in [23].

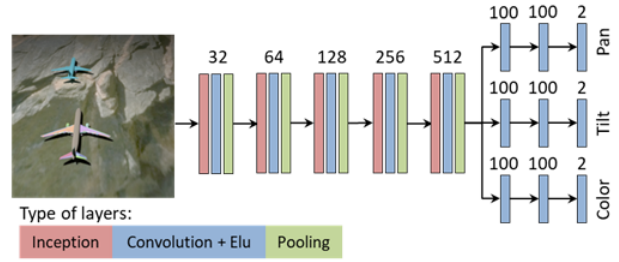


Figure 2. Deep Architecture. Inception module from [25]

Our deep architecture

We propose an inception-based encoder-decoder architecture to predict light parameters. In Figure 2 we give a scheme, where we can see that our encoder has five modules combining 3 types of layers: inception, convolution and pooling. The encoder input is the image that is transformed to a higher dimensional feature space, from which three decoders convert this embedding to a common feature space of pan, tilt and color of light source. Pan and tilt output predictions are given as functions of angle differences. We use the functions $\sin(\vartheta_c - \vartheta_l)$ and $\cos(\vartheta_c - \vartheta_l)$ to bound the pan output. Similarly, tilt prediction is represented as difference of angles $\sin(\varphi_c - \varphi_l)$ and $\cos(\varphi_c - \varphi_l)$. Finally color is predicted here as R, G and B values. We used the split inception module from [25], which replaces $n \times n$ convolution filters with $1 \times n$ and $n \times 1$ filter, to achieve faster convergence with overall less parameter. Our global loss function to estimate illumination parameters is based on three terms:

$$Loss(x, \hat{y}) = \alpha_1 L_{Pan}(x, \hat{y}) + \alpha_2 L_{Tilt}(x, \hat{y}) + \alpha_3 L_{Color}(x, \hat{y}) \quad (1)$$

where x is the input image, \hat{y} is a 7 dimensional vector giving the estimation of the scene light properties represented by x , α_i are the weights for the different loss terms defined for pan, tilt and color, and which are respectively given by:

$$\begin{aligned} L_{Pan}(x, \hat{y}) &= MSE((\hat{y}_1 - \sin(\vartheta_c^x - \vartheta_l^x)) + (\hat{y}_2 - \cos(\vartheta_c^x - \vartheta_l^x))) \\ L_{Tilt}(x, \hat{y}) &= MSE\{(\hat{y}_3 - \sin(\varphi_c^x - \varphi_l^x)) + (\hat{y}_4 - \cos(\varphi_c^x - \varphi_l^x))\} \\ L_{Color}(x, \hat{y}) &= \arccos((\hat{y}_5 \cdot x_{RGB}) / \|\hat{y}_5\| * \|x_{RGB}\|) \end{aligned}$$

L_{Pan} and L_{Tilt} are computed as the mean square error (MSE) between the estimations for pan, \hat{y}_1 and \hat{y}_2 , and for tilt, \hat{y}_3 and \hat{y}_4 , and a function of the difference between the camera and light positions for the ground-truth of x . The third loss term, L_{Color} , is the mean angular error between the estimated RGB values, \hat{y}_5 , and the color of the light for x image provided in the ground-truth.

This network has been trained using Adam optimizer [15], with initial learning rate 0.0002 which is decreased with factor of 0.1 on reaching plateau. Weights are initialized using He Normal[14]. All experiments in next sections were trained using a batch size of 16. In the following sections we show the results of several experiments to evaluate the architecture performance on different datasets and conditions.

Experiment 1: Synthetic dataset

In this first experiment we trained and tested the proposed architecture on two different datasets: SID1 (single object) and SID2 (multiple objects). The results are shown in Table 1, where we separately compute different angular errors. Direction error is given separately in the pan and tilt components, and the global angular error for direction estimation. We can see that all the estimations are improved when the network is trained on a more complex dataset, like SID2, where multiple objects light interactions provide richer shading cues. However, there is slight improvement for color estimation, performance is similar for both

datasets. In Figure 3 we show qualitative results on SID2 dataset. Images are ordered from smaller (left) to larger (right) direction estimation error.

Dataset	Pan	Tilt	Direction	Color
SID1	14.63	9.86	16.98	1.05
SID2	10.46	9.21	14.22	1.02

Table 1. Estimation Errors (in degrees) for light source direction and color with the proposed architecture trained on SID1 and SID2.

Intuitively as the light becomes more zenithal, the shadows shorten and cast shadows present more uncertainty to estimate light direction. We analysed the performance of the method at different tilt locations of the light source in the input image, from the ground (level 1: $[30^\circ, 50^\circ]$) up to the zenithal area (level 3: $[70^\circ, 90^\circ]$). This effect is confirmed in Table 2, where estimated errors in direction clearly increase from level 1 to level 3.

Tilt range	Pan	Tilt	Direction	Color
Level 1	4.92	5.14	7.57	1.04
Level 2	8.51	5.14	11.97	0.90
Level 3	22.36	18.33	28.33	1.00

Table 2. Estimation Errors (in degrees) for light direction and color at different tilt levels.

Similarly, we analysed the performance at different pan levels, each level covers 90° of pan area. Level 1 is when light comes from center front, level 2 from right, level 3 from back and level 4 is when light comes from left side. Tilt angle was kept between 30° and 70° to analyze pan error while minimizing zenithal tilt error effects. Table 3 shows results for this experiment, both direction and color error are consistent in all levels of pan.

Pan range	Pan	Tilt	Direction	Color
Level1	6.87	5.10	9.10	0.98
Level2	6.24	5.12	8.80	0.96
Level3	6.35	5.09	9.46	1.03
Level4	6.01	5.16	9.03	0.97

Table 3. Estimation Errors (in degrees) for light direction and color at different pan levels.

Experiment 2. Multi-illumination dataset

Once we have evaluated our method in synthetic images, we want to analyze whether it generalises for real images. We have tested our method on the *Multi-illuminant dataset* (MID) [18], that contains 1000 different indoor scenes, all of them containing a diffuse and a specular sphere at random locations. Light source is mounted above the camera and can be rotated at different predefined pan and tilt angles, creating different light conditions. The dataset provides the orientation of the light source for each acquired image, but since the light can bounce off the walls, the direction of the incident light on the scene is not defined by the light source angles and it needs to be recomputed.

We have defined a procedure to compute the incident light direction from the specular sphere present in all the scenes, whose highlights provide enough information to collect our GT data (tilt and pan angle between light and camera). The color of the light is obtained from the average color of the diffuse sphere. To obtain the light direction we used the ideas proposed by [22] where they assume that the angle of incident light is equal to the angle of outgoing light at the specular highlight on a spherical ball. We use a reference image in each scene where light and

camera both are pointed in the same direction towards the center of the scene. We also assumed that the light is mounted at 10° height with respect to scene center. The angles obtained from this reference images allow to correct the angle displacement due to the sphere position shifting inside the image on the rest of scene images.

Dataset (Error)	Pan	Tilt	Direction	Color
Masked MID (Mean)	21.38	10.14	22.72	0.63
Masked MID (Median)	13.74	7.64	17.80	0.40
UnMasked MID (Mean)	14.28	6.96	15.44	0.36
UnMasked MID (Median)	8.20	4.83	10.83	0.24

Table 4. Estimation errors in degrees on two versions of MID dataset (with Masked or UnMasked spheres).

Starting from the network trained on SID2 it was fine tuned on this dataset under two different conditions: a) keeping the reference spheres in the image, and b) masking them. Although specular spheres are not present in real images, the first configuration should provide an upper bound of our method performance on wild images. Table 4 shows the results on this experiment. As expected, network performance is much higher when complete images are used as inputs. We can also observe that results on color estimation are better than on SID2, mainly due to the stability of single white light source in the dataset. To analyze the results removing the influence of the outliers, we also reported median error on this dataset. Results are as good as the ones obtained only using synthetic images on the upper bound. Qualitative results are provided in Figure 4. Top row depicts the original image, second row are the spheres generated from the GT information, and the third row shows the synthetic spheres generated with the obtained prediction.

Finally, we perform a last experiment on this dataset by dividing the test set in two: (a) images with incident light from the front, and (b) from the back. Table 5 also shows the errors computed for these two sets. Both color and direction errors are higher when the light comes from the back of the scene and a big area of the image becomes saturated. We want to note here that the GT we created present a low accuracy for the subset of images with back light sources. This is due to the inherent uncertainty derived from what can be inferred from spheres illuminated from the back. Therefore, this MID dataset division is highly recommended to analyse results derived from this GT.

Light position	Pan	Tilt	Direction	Color
Front	11.51	6.33	13.09	0.34
Back	32.90	11.21	31.23	0.52

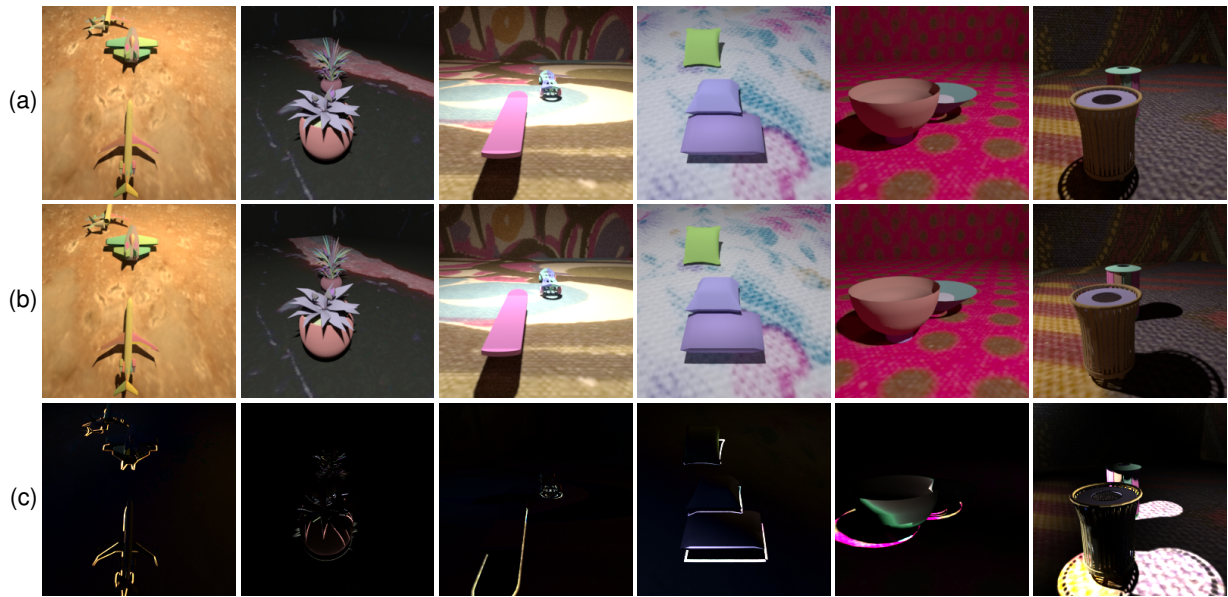
Table 5. Estimation errors in degrees dividing MID dataset in front and back light.

Experiment 3. Natural images

Previous experiments show the performance of our method on synthetic and real indoor images. Here, we show a few qualitative results on real outdoor images. In Figure 5 we show some examples with strong outdoor cast shadows, in order to visually evaluate the prediction we depict a synthetic pole at the left top corner. In these examples camera is assumed to be at 45° tilt from ground. Left side four images are from SBU shadow dataset [26] and the two on the right have been captured with a mobile device.

Conclusions

In this work we have proved the plausibility of using a simple deep architecture to estimate physical light properties of a scene from a single image. The proposed approach is based on



<i>Direction</i>	0.57	1.58	4.3	8.55	8.81	30.0
<i>Color</i>	1.27	0.78	1.64	1.8	0.22	0.24

Figure 3. Direction and Color estimation examples on SID2 dataset: (a) Original images, (b) Generated images with estimated light properties, (c) RGB Image subtraction between (a) and (b). Bottom rows are the corresponding computed errors for direction and color in degrees, ordered from smaller (left) to larger (right) direction estimation error.



<i>Direction</i>	3.16	5.60	20.12	25.05
<i>Color</i>	0.16	0.21	1.30	0.35

Figure 4. Direction and Color estimation examples on Multi-illumination dataset: (a) Original images, (b) Ground-truth plotted on corresponding spheres, (c) Estimations provided by our proposed architecture. Bottom rows are computed errors for direction and color in degrees.



Figure 5. Examples of light direction estimation on natural images. Predicted direction is plotted top left in each image.

training a deep regression architecture on a large synthetic and diversified dataset. We show that the obtained regressor can generalize to real images and can be used as a preliminary step for further complex tasks.

References

- [1] K. Agusanto, L. Li, Z. Chuangui, and W.S. Ng. Photorealistic rendering for augmented reality using environment illumination. *ISMAR*, pages 208–216, 2003.
- [2] I. Arief, S. McCallum, and J.Y. Hardeberg. Realtime estimation of illumination direction for augmented reality on mobile devices. In *CIC*, volume 2012, pages 111–116, 2012.

- [3] A.X. Chang, T.A. Funkhouser, L.J. Guibas, P. Hanrahan, Q. Huang, Z. Li, S. Savarese, M. Savva, S. Song, H. Su, J. Xiao, L. Yi, and F. Yu. Shapenet: An information-rich 3d model repository. *CoRR*, abs/1512.03012, 2015.
- [4] D. Cheng, D.K. Prasad, and M.S. Brown. Illuminant estimation for color constancy: why spatial-domain methods work and the role of the color distribution. *JOSA A*, 31(5):1049–1058, 2014.
- [5] P. Das, A.S. Baslamisli, Y. Liu, S. Karaoglu, and T. Gevers. Color constancy by gans: An experimental survey, 2018.
- [6] P.E. Debevec. Rendering synthetic objects into real scenes: bridging traditional and image-based graphics with global illumination and high dynamic range photography. In *SIGGRAPH*, 1998.
- [7] G.D. Finlayson, S.D. Hordley, C. Lu, and M.S. Drew. On the removal of shadows from images. *PAMI*, 28(1):59–68, 2006.
- [8] D.H. Foster. Color constancy. *Vision Research*, 51(7):674 – 700, 2011.
- [9] M. Gardner, Y.H. Geoffroy, K. Sunkavalli, C. Gagné, and J. Lalonde. Deep parametric indoor lighting estimation. In *ICCV*, pages 7175–7183, 2019.
- [10] M. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J. Lalonde. Learning to predict indoor illumination from a single image. *arXiv preprint arXiv:1704.00090*, 2017.
- [11] Y.H. Geoffroy, K. Sunkavalli, S. Hadap, E. Gambaretto, and J. Lalonde. Deep outdoor illumination estimation. In *CVPR*, pages 7312–7321, 2017.
- [12] A. Gijsenij, T. Gevers, and J.V. Weijer. Computational color constancy: Survey and experiments. *PAMI*, 20(9):2475–2489, 2011.
- [13] J.F. Lalonde, A.A. Efros, and S.G. Narasimhan. Estimating natural illumination from a single outdoor image. *ICCV*, pages 183–190, 2009.
- [14] H. Kaiming, Z. Xiangyu, R. Shaoqing, and S. Jian. Deep residual learning for image recognition. *CVPR*, pages 770–778, 2016.
- [15] D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.
- [16] Z. Lou, T. Gevers, N. Hu, and M.P. Lucassen. Color constancy by deep learning. In *BMVC*, pages 76.1–76.12, September 2015.
- [17] D. Mandl, K.M. Yi, P. Mohr, P.M. Roth, P. Fua, V. Lepetit, D. Schmalstieg, and D. Kalkofen. Learning lightprobes for mixed reality illumination. *ISMAR*, pages 82–89, 2017.
- [18] L. Murmann, M. Gharbi, M. Aittala, and F. Durand. A dataset of multi-illumination images in the wild. In *ICCV*, pages 4080–4089, 2019.
- [19] A. Panagopoulos, D. Samaras, and N. Paragios. Robust shadow and illumination estimation using a mixture model. In *CVPR*, pages 651–658. IEEE, 2009.
- [20] A. Panagopoulos, C. Wang, D. Samaras, and N. Paragios. Illumination estimation and cast shadow detection through a higher-order graphical model. *CVPR 2011*, pages 673–680, 2011.
- [21] D. Samaras and D. Metaxas. Incorporating illumination constraints in deformable models for shape from shading and light direction estimation. *PAMI*, 25(2):247–264, 2003.
- [22] D. Schnieders. Light source estimation from spherical reflections. 2011.
- [23] H.A. Sial, R. Baldrich, and M. Vanrell. Deep intrinsic decomposition trained on surreal scenes yet with realistic light effects. *JOSA A*, 37(1):1–15, 2020.
- [24] T. Sun, J.T. Barron, Y. Tsai, Z. Xu, X. Yu, G. Fyffe, C. Rhemann, J. Busch, P.E. Debevec, and R. Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 38:79:1–79:12, 2019.
- [25] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016.
- [26] Tomás F. Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling, editors, *Computer Vision – ECCV 2016*, 2016.
- [27] S. Yan, F. Peng, H. Tan, S. Lai, and M. Zhang. Multiple illumination estimation with end-to-end network. *ICIVC*, pages 642–647, 2018.