Spatiotemporal Outdoor Lighting Aggregation on Image Sequences

Haebom Lee[1,2](✉), Robert Herzog[1], Jan Rexilius[3], and Carsten Rother[2]

[1] Computer Vision Research Lab, Robert Bosch GmbH, Hildesheim, Germany
[2] Computer Vision and Learning Lab, Heidelberg University, Heidelberg, Germany
[3] Campus Minden, Bielefeld University of Applied Sciences, Minden, Germany

Abstract. In this work, we focus on outdoor lighting estimation by aggregating individual noisy estimates from images, exploiting the rich image information from wide-angle cameras and/or temporal image sequences. Photographs inherently encode information about the scene's lighting in the form of shading and shadows. Whereas computer graphic (CG) methods target accurately reproducing the image formation process knowing the exact lighting in the scene, the inverse rendering is an ill-posed problem attempting to estimate the geometry, material, and lighting behind a recorded 2D picture. Recent work based on deep neural networks has shown promising results for single image lighting estimation despite its difficulty. However, the main challenge remains on the stability of measurements. We tackle this problem by combining lighting estimates from many image views sampled in the angular and temporal domain of an image sequence. Thereby, we make efficient use of the camera calibration and camera ego-motion estimation to globally register the individual estimates and apply outlier removal and filtering algorithms. Our method not only improves the stability for rendering applications like virtual object augmentation but also shows higher accuracy for single image based lighting estimation compared to the state-of-the-art.

Keywords: Lighting estimation · Spatio-temporal filtering · Virtual object augmentation

1 Introduction

Deep learning has shown its potential in estimating hidden information like depth from monocular images [4] by only exploiting learned priors. Accordingly, it has also been applied for the task of lighting estimation. The shading in a photograph captures the incident lighting (irradiance) on a surface point. It depends not only on the local surface geometry and material but also on the global (possibly occluded) lighting in a mostly unknown 3D scene. Different configurations of

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/978-3-030-92659-5_22.

C. Bauckhage et al. (Eds.): DAGM GCPR 2021, LNCS 13024, pp. 343–357, 2021.
https://doi.org/10.1007/978-3-030-92659-5_22

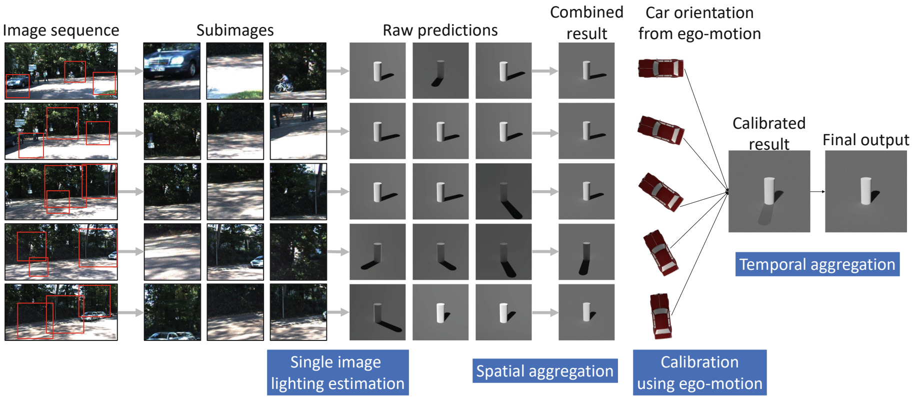


Fig. 1. Spatio-temporal outdoor lighting aggregation on an image sequence: individual estimates from each generated subimage are combined in the spatial aggregation step. Spatial aggregation results for each image in the sequence are then calibrated using camera ego-motion data and further refined in the temporal aggregation step to generate the final lighting estimate for the sequence.

material, geometry, and lighting parameters may lead to the same pixel color, which creates an ill-posed optimization problem without additional constraints. Hence, blindly estimating the lighting conditions is notoriously difficult, and we restrict ourselves to outdoor scenes considering only environment lighting where the incident lighting is defined to be spatially invariant.

Estimating environment lighting can be regarded as the first step towards holistic scene understanding and enables several applications [1,11,22,28]. It is essential for augmented reality (seamlessly rendering virtual objects into real background images) because photo-realistically inserting virtual objects in real images requires knowing not just the 3D geometry and camera calibration but also the lighting. The human eye quickly perceives wrong lighting and shadows as unrealistic, and it has also been shown [27] that shadows are essential for depth-from-mono estimation using convolutional neural networks.

There have been numerous studies on estimating the lighting from image data. Those methods mostly focus on estimating sky map textures [5] or locating the sun position from a single RGB image [6,8,34], calculating sun trajectories from time-lapse videos [1,19], or utilizing material information to conjecture the positions of multiple light sources [29].

In this paper, we propose a method to robustly estimate the global environment's sun direction by exploiting temporal and spatial coherency in outdoor lighting. The image cues for resolving the lighting in a scene appear sparsely (e.g., shadows, highlights, etc.) or very subtle and noisy (e.g., color gradients, temperature, etc.) and not all images provide the same quality of information for revealing the lighting parameters. For example, consider an image view completely covered in shadow. Hence, the predictions for the lighting on individual images of a sequence are affected by a large amount of noise and many outliers. To alleviate this issue we propose to sample many sub-views of an image sequence essentially sampling in the angular and temporal domain. This approach has two advantages. First, we effectively filter noise and detect outliers, and second, our neural network-based lighting estimator becomes invariant to the imaging parameters like size, aspect ratio, and camera focal length and can explore details in the high-resolution image content. To this end, the contributions of this paper are:

1. A single image based sunlight estimation using a deep artificial neural network that is on par or better than the current state-of-the-art,
2. A two-stage post-processing approach for spatial and temporal filtering with outlier detection that fully exploits the information from calibrated image sequences to overcome noisy, outlier-sensitive estimation methods.

2 Related Work

Outdoor lighting condition estimation has been studied in numerous ways because of its importance in computer graphics and computer vision applications [12,20]. Related techniques can be categorized into two parts, one that analyzes a single image [5,9,15,21] and the other that utilizes a sequence of images [1,16,19,22]. For example, the outdoor illumination estimation method presented in [23] belongs to the latter as the authors estimated the sun trajectory and its varying intensity from a sequence of images. Under the assumption that a static 3D model of the scene is available, they designed a rendering equation-based [10] optimization problem to determine the continuous change of the lighting parameters. On the other hand, Hold-Geoffroy et al. [6] proposed a method that estimates outdoor illumination from a single low dynamic range image using a convolutional neural network [14]. The network was able to classify the sun location on 160 evenly distributed positions on the hemisphere and estimated other parameters such as sky turbidity, exposure, and camera parameters.

Analyzing outdoor lighting conditions is further developed in [34] where they incorporated a more delicate illumination model [16]. The predicted parameters were numerically compared with the ground truth values and examined rather qualitatively by utilizing the render loss. Jin et al. [8] and Zhang et al. [35] also proposed single image based lighting estimation methods. While their predecessors [6,34] generated a probability distribution of the sun position on the discretized hemisphere, the sun position parameters were directly regressed from their networks. Recently, Zhu et al. [36] combined lighting estimation with intrinsic image decomposition. Although they achieved a noticeable outcome in the sun position estimation on a synthetic dataset, we were unable to compare it with ours due to the difference in the datasets.

The aforementioned lighting estimation techniques based on a single image often suffer from insufficient cues to determine the lighting condition, for example, when the given image is in a complete shadow. Therefore, several attempts were made to increase the accuracy and robustness by taking the temporal domain into account [1,16,22]. The method introduced in [19] extracts a set of features from each image frame and utilizes it to estimate the relative changes of the lighting parameters in an image sequence. Their method is capable of handling a moving camera and generating temporally coherent augmentations. However, the estimation process utilized only two consecutive frames and assumed that the sun position is given in the form of GPS coordinates and timestamps [25].

Lighting condition estimation is also crucial for augmented reality where virtual objects become more realistic when rendered into the background image using the correct lighting. Lu et al. [20], for instance, estimated a directional light vector from shadow regions and the corresponding objects in the scene to achieve realistic occlusion with augmented objects. The performance of the estimation depends solely on the shadow region segmentation and finding related items. Therefore, the method may struggle if a shadow casting object is not visible in the image. Madsen and Lal [22] utilize a stereo camera to extend [23] further. Using the sun position calculated from GPS coordinates and timestamps, they estimated the variances of the sky and the sun over an image sequence. The estimation is then combined with a shadow detection algorithm to generate plausible augmented scenes with proper shading and shadows.

Recently, there have been several attempts utilizing auxiliary information to estimate the lighting condition [11,33]. Such information may result in better performance but only with a trade-off in generality. Kán and Kaufmann [11] proposed a single RGB-D image-based lighting estimation method for augmented reality applications. They utilized synthetically generated scenes for training a deep neural network, which outputs the dominant light source's angular coordinates in the scene. Outlier removal and temporal smoothing processes were applied to make the method temporally consistent. However, their technique was demonstrated only on fixed viewpoint scenes. Our method, on the other hand, improves its estimation by aggregating observations from different viewpoints. We illustrate the consistency gained from our novel design by augmenting virtual objects in consecutive frames.

3 Proposed Method

We take advantage of different aspects of previous work and refine them into our integrated model. As illustrated in Fig. 1, our model is composed of four subprocesses. We first randomly generate several small subimages from an input image and upsample them to a fixed size. Since modern cameras are capable of capturing fine details of a scene, we found that lighting condition estimation can be done on a small part of an image. These spatial samples obtained from one image all share the same lighting condition and therefore yield more robustness compared to a single image view. Then, we train our lighting estimation network on each sample to obtain the global lighting for a given input image.

After the network estimates the lighting conditions for the spatial samples, we perform a spatial aggregation step to get a stable prediction for each image. Note that the estimate for each frame is based on its own camera coordinate system. Our third step is to unify the individual predictions into one global coordinate system using the camera ego-motion. Lastly, the calibrated estimates are combined in the temporal aggregation step. The assumption behind our approach is that distant sun-environment lighting is invariant to the location the picture was taken and that the variation in lighting direction is negligible for short videos. Through the following sections, we introduce the details of each submodule.

3.1 Lighting Estimation

There have been several sun and sky models to parameterize outdoor lighting conditions [7,16]. Although those methods are potentially useful to estimate complex lighting models consistently, in this work we focus only on the most critical lighting parameter: the sun direction. The rationale behind this is that ground-truth training data can easily be generated for video sequences with GPS and timestamp information (e.g., KITTI dataset [3]). Therefore, the lighting estimation network's output is a 3D unit vector $\vec{v}_{pred}$ pointing to the sun's location in the camera coordinate system.

Unlike our predecessors [6,34], we design our network as a direct regression model to overcome the need for a sensitive discretization of the hemisphere. The recent work of Jin et al. [8] presented a regression network estimating the sun direction in spherical coordinates (altitude and azimuth). Our method, however, estimates the lighting direction using Cartesian coordinates and does not suffer from singularities in the spherical parametrization and the ambiguity that comes from the cyclic nature of the spherical coordinates.

Since we train our network in a supervised manner, the loss function is defined to compare the estimated sun direction with the ground truth $\vec{v}_{gt}$:

$$L_{cosine} = 1 - \vec{v}_{gt} \cdot \vec{v}_{pred}/||\vec{v}_{pred}||, \tag{1}$$

with the two adjacent unit vectors having their inner product close to 1. To avoid the uncertainty that comes from the vectors pointing the same direction with different lengths, we apply another constraint to the loss function:

$$L_{norm} = (1 - ||\vec{v}_{pred}||)^2. \tag{2}$$

The last term of the loss function ensures that the estimated sun direction resides in the upper hemisphere because we assume the sun is the primary light source in the given scene:

$$L_{hemi} = max(0, -z_{pred}), \tag{3}$$

where z_{pred} is the third component of $\vec{v}_{pred}$, indicating the altitude of the sun. The final loss function is simply the sum of all terms as they share a similar range of values:

$$L_{light} = L_{cosine} + L_{norm} + L_{hemi}. \tag{4}$$

3.2 Spatial Aggregation

Using our lighting estimator, we gather several lighting condition estimates from different regions of the image. Some of those estimates may contain larger errors due to insufficient information in the given region to predict the lighting condition. We refer to such estimates as outliers. Our method's virtue is to exclude anomalies that commonly occur in single image-based lighting estimation techniques and deduce the best matching model that can explain the inliers.

Among various outlier removal algorithms, we employ the isolation forest (iForest) algorithm [18]. The technique is specifically optimized to isolate anomalies instead of building a model of inliers and eliminate samples not complying with it. In essence, the iForest algorithm recursively and randomly splits the feature space into binary decision trees (hence forming a forest). Since the outliers are outside of a potential inlier cluster, a sample is classified as an outlier if the sample's average path length is shorter than a threshold (*contamination ratio* [24]). We determine this value empirically and use it throughout all results.

On the remaining inliers, we apply the *mean shift algorithm* [2] to conjecture the most feasible lighting parameters. Unlike naive averaging over all inliers, this process further refines the lighting estimate by iteratively climbing to the maximum density in the distribution. Another experimentally discoverable parameter *bandwidth* determines the size of the Gaussian kernel to measure the samples' local density gradient. In the proposed method, we set the bandwidth as the median of all samples' pairwise distances. By moving the data points iteratively towards the closest peak in the density distribution, the algorithm locates the highest density within a cluster, our spatial aggregation result. We compare various aggregation methods in the ablation study in Sect. 4.4.

3.3 Calibration

Since our primary goal is to assess the sun direction for an input video, we perform a calibration step to align the estimates because the sun direction determined from each image in a sequence is in its own local camera coordinate system. The camera ego-motion data is necessary to transform the estimated sun direction vectors into the world coordinate system. We assume the noise and drift in the ego-motion estimation is small relative to the lighting estimation. Hence, we employ a state-of-the-art structure-from-motion (SfM) technique such as [26] to estimate the ego-motion from an image sequence. Then there exists a camera rotation matrix R_f for each frame f and the resulting calibrated vector $\hat{\vec{v}}_{pred}$ is computed as $R_f^{-1} \cdot \vec{v}_{pred}$.

3.4 Temporal Aggregation

Having the temporal estimates aligned in the same global coordinate system, we consider them as independent observations of the same lighting condition in the temporal domain. Although the lighting estimates from our regression network are not necessarily independent for consecutive video frames, natural image sequences, as shown empirically in our experiments, reveal a large degree

Table 1. Number of data and subimages for training and test

Dataset		SUN360	KITTI
Training	Data	16 891	3630
	Subimg	135 128	116 160
Test	Data	1688	281
	Subimg	108 032	17 984

of independent noise in the regression results, which is however polluted with a non-neglectable amount of outliers. Consequently, we apply a similar aggregation strategy as in the spatial domain also for the temporal domain. Therefore, the final output of our pipeline, the lighting condition for the given image sequence, is the mean shift algorithm's result on the inliers from all frames of the entire image sequence.

4 Experiments

4.1 Datasets

One of the common datasets considered in the outdoor lighting estimation methods is the SUN360 dataset [31]. Several previous methods utilized it in its original panorama form or as subimages by generating synthetic perspective images [6]. We follow the latter approach since we train our network using square images. We first divide 20267 panorama images into the training, validation, and test sets with a 10:1:1 ratio. For the training and the validation sets, 8 subimages from each panorama are taken by evenly dividing the azimuth range. To increase the diversity, 64 subimages with random azimuth values are generated from each panorama in the test set. Note that we introduce small random offsets on the camera elevation with respect to the horizon in $[-10°, 10°]$ and randomly select a camera field of view within a range $[50°, 80°]$. The generated images are resized to 256×256. In this way, we produced 135128, 13504, and 108032 subimages from 16891, 1688, and 1688 panoramas for the training, validation, and test sets, respectively. The ground truth labeling was given by the authors of [34].

The well-known KITTI dataset [3] has also attracted our attention. Since the dataset is composed of several rectified driving image sequences and provides the information required for calculating the ground truth sun directions [25], we utilize it for both training and test. Specifically, since the raw data was recorded at 10 Hz, we collect every 10^{th} image to avoid severe repetition and split off five randomly chosen driving scenes for validation and test set. The resulting training set is composed of 3630 images. If we train our network using only one crop for each KITTI image, the network is likely to be biased to the SUN360 dataset due to the heavy imbalance in the amount of data. To match the number of samples, we crop 32 subimages from one image by varying the cropping location and the crop size. Each image in the test set is again cropped into 64 subimages and the cropped images are also resized to 256×256. In total, we train our network on about 250000 images. The exact numbers of samples are presented in Table 1 and Fig. 2 illustrates examples of the two datasets.

4.2 Implementation Details

Our lighting estimation model is a regression network with convolution layers. It accepts an RGB image of size 256 × 256 and outputs the sun direction estimate. We borrow the core structure from ResNeXt [32] and carefully determine the number of blocks, groups, and filters as well as the sizes of filters under extensive experiments. As illustrated in Fig. 3, the model is roughly composed of 8 bottleneck blocks, each of which is followed by a convolutional block attention module [30]. In this way, our network is capable of focusing on important spatial and channel features while acquiring resilience from vanishing or exploding gradients by using the shortcut connections. A global average pooling layer is adopted to connect the convolution network and the output layer and serves as a tool to mitigate possible overfitting [17]. The dense layer at the end then refines the encoded values into the sun direction estimate.

We train our model and test its performance on the SUN360 and the KITTI datasets (see Table 1). In detail, we empirically trained our lighting estimation network for 18 epochs using early stopping. The training was initiated with the Adam optimizer [13] using a learning rate of 1×10^{-4} and the batch size was 64. It took 12 h on a single Nvidia RTX 2080 Ti GPU. Prediction on a single image takes 42 ms. Our single image lighting estimation and spatial aggregation modules are examined upon 108 032 unobserved SUN360 crops generated from 1688 panoramas. The whole pipeline including the calibration and temporal aggregation modules is analyzed on five unseen KITTI sequences composed of 281 images.

4.3 Results

We evaluate the angular errors of the spatially aggregated sun direction estimates on the SUN360 test set. At first, single image lighting estimation results

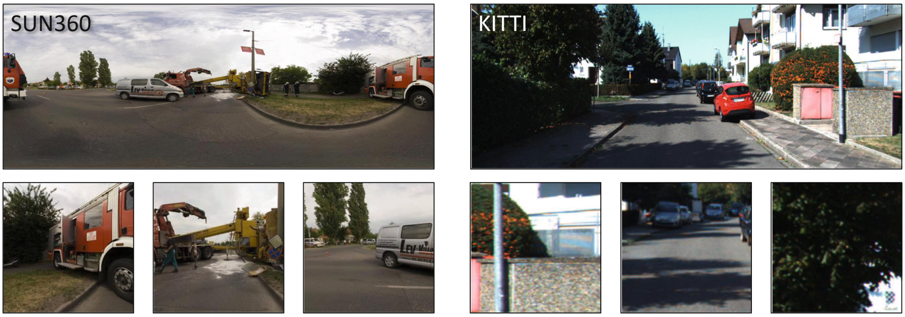


Fig. 2. Examples of the two datasets [3,31]. From the original image (*top*), we generate random subimages (*bottom*).

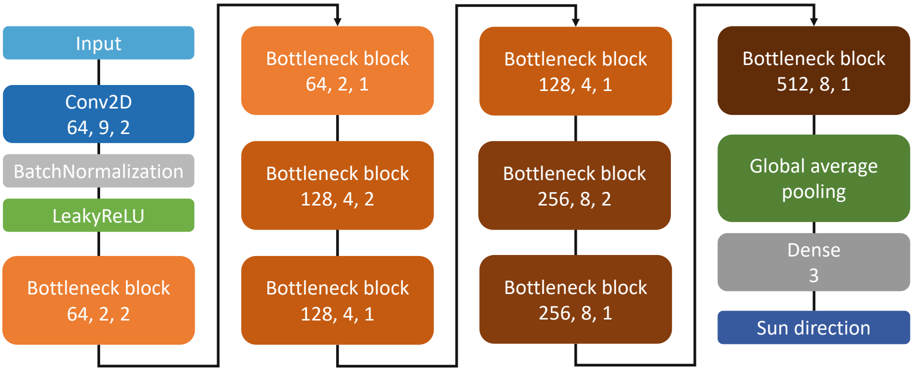


Fig. 3. The proposed lighting estimation network. The numbers on the *Conv2D* layer indicate the number of filters, the filter size, and the stride, whereas the numbers on each *Bottleneck block* depict the number of 3 × 3 filters, the cardinality, and the stride. A *Bottleneck block* is implemented following the structure proposed in [32] except for a convolutional block attention module [30] attached at the end of each block.

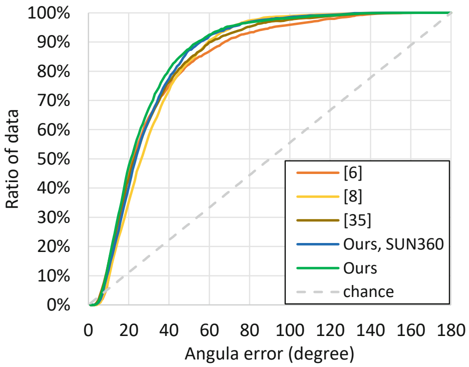


Fig. 4. The cumulative angular error for spatially aggregated sun direction estimates on the SUN360 test set. *Ours, SUN360* indicates our results when the network was only trained with the SUN360 dataset.

are gathered using [6,8,35], and our method. Then we compensate the camera angles and apply our spatial aggregation step on the subimages to acquire the spatially combined estimate for each panorama. The explicit spatial aggregation step involves two additional hyperparameters: the contamination ratio and the mean-shift kernel width. We found those parameters to be insensitive to different data sets and kept the same values in all our experiments. The *contamination ratio* is set to 0.5 because we assume the estimations with angular errors larger than an octant (22.5°) as outliers, which is roughly 50% of the data for our method when observing Fig. 7. As a result, we apply the mean shift algorithm on 50% potential inliers among the total observations.

Table 2. Angular errors of each aggregation step (from left to right: single image (baseline), spatial aggregation, spatio-temporal aggregation). Sequences correspond to Fig. 5.

Sequence	Single	Spatial	Spatiotemporal
(a)	13.43	6.76	3.54
(b)	26.06	7.81	6.87
(c)	34.68	24.83	13.17
(d)	23.03	10.04	3.27

Figure 4 illustrates the cumulative angular errors of the four methods. Since the previous methods were trained with only the SUN360 training set, due to the characteristics of their networks (requiring ground truth exposure and turbidity information which are lacked in the KITTI dataset), we also report our method's performance when it was trained only on SUN360 (see *Ours, SUN360* in Fig. 4). Our method performs better than the previous techniques even with the same training set. The detailed quantitative comparison is presented in Fig. 7. Note that all methods are trained and tested with subimages instead of full images.

For the KITTI dataset, we can further extend the lighting estimation to the temporal domain. Although the dataset provides the ground truth ego-motion, we calculated it using [26] to generalize our approach. The mean angular error of the estimated camera rotation using the default parameters was 1.01° over the five test sequences. We plotted the sun direction estimates of each step in our pipeline for four (out of five) test sequences in Fig. 5. Note that in the plots all predictions are registered to a common coordinate frame using the estimated camera ego-motion. Individual estimates of the subimages are shown with gray dots. Our spatial aggregation process refines the noisy observations using outlier removal and mean shift (black dots). Those estimates for each frame in a sequence are finally combined in the temporal aggregation step (denoted with the green dot). The ground truth direction is indicated by the red dot. Using the spatio-temporal filtering, the mean angular error over the five test sequences recorded 7.68°, which is a reduction of 69.94% (25.56° for single image based estimation). A quantitative evaluation of the performance gain for each aggregation step is presented in Table 2.

Our model's stability is better understood with a virtual object augmentation application as shown in Fig. 6. Note that other lighting parameters, such as the sun's intensity are manually determined. When the lighting conditions are estimated from only a single image on each frame, the virtual objects' shadows are fluctuating compared to the ground truth results. The artifact is less visible on our spatial aggregation results and entirely removed after applying the spatio-temporally aggregated lighting condition.

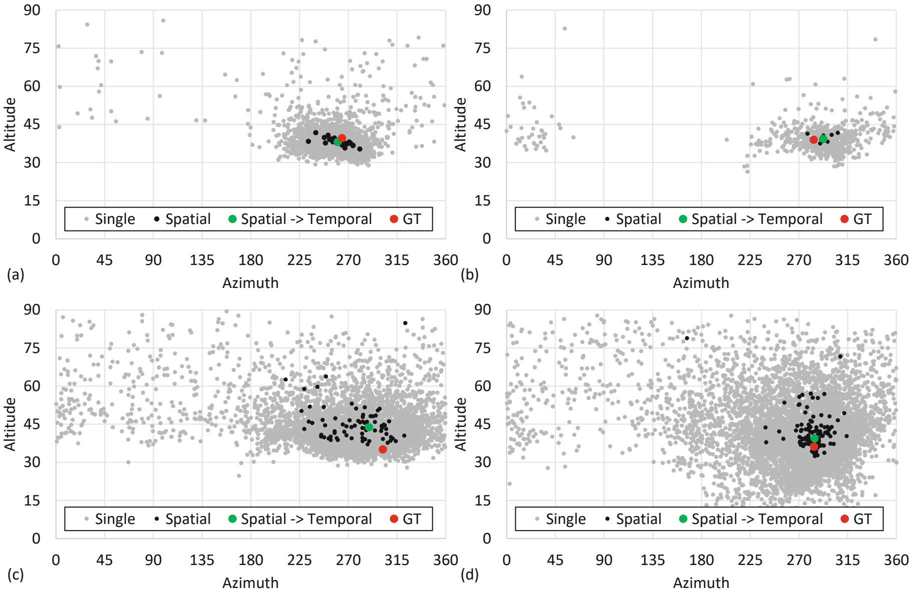


Fig. 5. Scatter plots representing sun direction estimates of individual subimages and the results of two aggregation steps. Each graph corresponds to an image sequence in the KITTI test set. Despite numerous outliers in the raw observations (the gray dots), our two-step aggregation determines the video's lighting condition with small margins to the ground truth sun direction (the black dots for spatial aggregation and the green dot for spatio-temporal aggregation). Angular errors for our spatio-temporal filtering results are (a) 3.54 (b) 6.87 (c) 13.17 and (d) 3.27°. (Color figure online)

4.4 Ablation Study

The performance gain of the spatial aggregation process is thoroughly analyzed by breaking down the individual filtering steps on the SUN360 test set. Figure 7 shows the cumulative angular error for the raw observations and compares the four lighting estimation methods with four different aggregation strategies:

- *Single*: unprocessed individual observations,
- *Mean all*: mean of all estimates from each panorama,
- *Mean inliers*: mean of inlier estimates,
- *Meanshift*: mean shift result of inlier estimates.

As illustrated in Fig. 7, the average angular error of each method is decreased by at most 10° after applying the proposed spatial aggregation. This result demonstrates our method's generality, showing that it can increase the accuracy of any lighting estimation method. We observe a slight increase in the average error for the *Mean all* metric due to the outlier observations. A similar analysis

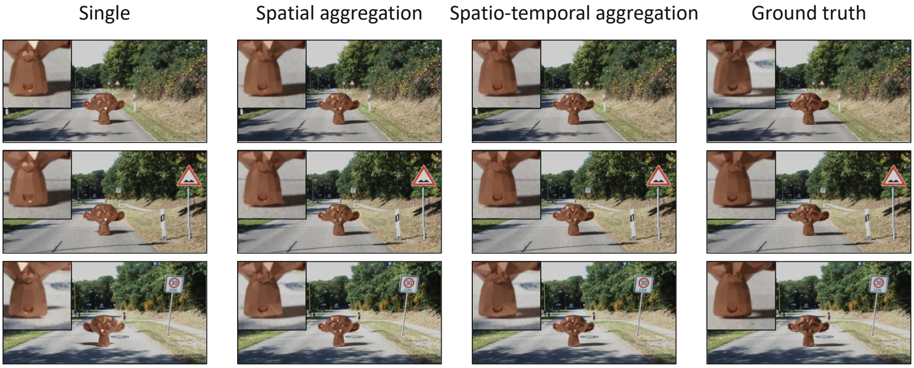
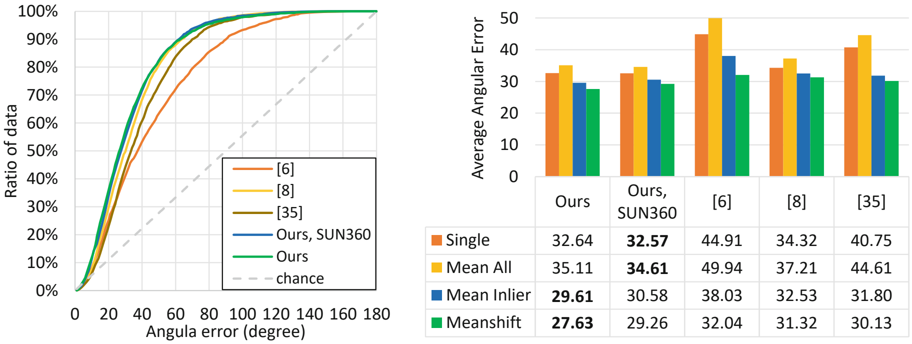


Fig. 6. Demonstration of a virtual augmentation application. Fluctuations in the shadow of the augmented object decrease as the estimates are refined through our pipeline. After applying the spatio-temporal filtering, the results are fully stabilized and almost indistinguishable from the ground truth. Please also refer to the augmented video in the supplementary material.



	Ours	Ours, SUN360	[6]	[8]	[35]
🟧 Single	32.64	**32.57**	44.91	34.32	40.75
🟨 Mean All	35.11	**34.61**	49.94	37.21	44.61
🟦 Mean Inlier	**29.61**	30.58	38.03	32.53	31.80
🟩 Meanshift	**27.63**	29.26	32.04	31.32	30.13

Fig. 7. (*left*) The cumulative angular error for the *single* estimates on the SUN360 test set. (*right*) Comparing average angular error for three methods with different spatial aggregation strategies. Our method achieved the best result when the mean shift is applied to the inliers. We outperform previous methods even without the KITTI dataset.

is done for the KITTI dataset with only our method. The cumulative angular error graphs for the four steps are presented in Fig. 8.

Our pipeline was also tested as an end-to-end model, but it failed to show comparable performance. We provide the details of this experiment as well as additional studies with different combinations of loss functions in the supplementary material.

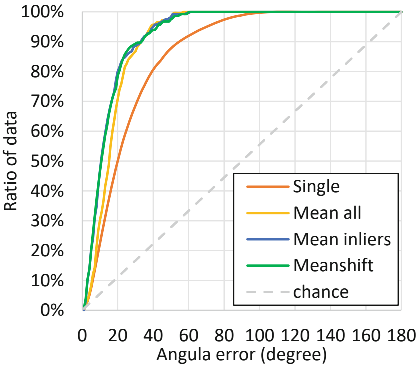


Fig. 8. The cumulative angular error on the KITTI test set with different spatial aggregation strategies. The best result is recorded when the mean shift result of the inlier estimates is utilized.

5 Conclusion

In this paper, we proposed a single image lighting estimation method and showed how its performance can be improved using spatial and temporal aggregation. Our method achieved state-of-the-art performance on outdoor lighting estimation for a given image sequence. We utilized 360° panoramas and wide view images in our work, but our spatial aggregation can be also applied to any image containing enough details. To this end, our spatio-temporal aggregation can be extended to different methods of gathering globally shared scene information.

Although we demonstrated noticeable outcomes in augmented reality applications, intriguing future research topics are remaining. We plan to extend our model to examine other factors such as cloudiness or exposure as it helps to accomplish diverse targets, including photorealistic virtual object augmentation over an image sequence. With such augmented datasets, we could enhance the performance of other deep learning techniques.

References

1. Balcı, H., Güdükbay, U.: Sun position estimation and tracking for virtual object placement in time-lapse videos. SIViP **11**(5), 817–824 (2017)
2. Comaniciu, D., Meer, P.: Mean shift: a robust approach toward feature space analysis. IEEE Trans. Pattern Anal. Mach. Intell. **24**(5), 603–619 (2002)
3. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? The KITTI vision benchmark suite. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3354–3361. IEEE (2012)
4. Godard, C., Mac Aodha, O., Firman, M., Brostow, G.J.: Digging into self-supervised monocular depth estimation. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3828–3838 (2019)

5. Hold-Geoffroy, Y., Athawale, A., Lalonde, J.F.: Deep sky modeling for single image outdoor lighting estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6927–6935 (2019)
6. Hold-Geoffroy, Y., Sunkavalli, K., Hadap, S., Gambaretto, E., Lalonde, J.F.: Deep outdoor illumination estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7312–7321 (2017)
7. Hosek, L., Wilkie, A.: An analytic model for full spectral sky-dome radiance. ACM Trans. Graph. (TOG) **31**(4), 1–9 (2012)
8. Jin, X., et al.: Sun-sky model estimation from outdoor images. J. Ambient Intell. Hum. Comput. 1–12 (2020). https://doi.org/10.1007/s12652-020-02367-3
9. Jin, X., et al.: Sun orientation estimation from a single image using short-cuts in DCNN. Optics Laser Technol. **110**, 191–195 (2019)
10. Kajiya, J.T.: The rendering equation. In: Proceedings of the 13th Annual Conference on Computer Graphics and Interactive Techniques, pp. 143–150 (1986)
11. Kán, P., Kaufmann, H.: Deeplight: light source estimation for augmented reality using deep learning. Vis. Comput. **35**(6–8), 873–883 (2019)
12. Karsch, K., Hedau, V., Forsyth, D., Hoiem, D.: Rendering synthetic objects into legacy photographs. ACM Trans. Graph. (TOG) **30**(6), 1–12 (2011)
13. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
15. Lalonde, J.F., Efros, A.A., Narasimhan, S.G.: Estimating the natural illumination conditions from a single outdoor image. Int. J. Comput. Vis. **98**(2), 123–145 (2012)
16. Lalonde, J.F., Matthews, I.: Lighting estimation in outdoor image collections. In: 2014 2nd International Conference on 3D Vision, vol. 1, pp. 131–138. IEEE (2014)
17. Lin, M., Chen, Q., Yan, S.: Network in network. arXiv preprint arXiv:1312.4400 (2013)
18. Liu, F.T., Ting, K.M., Zhou, Z.H.: Isolation forest. In: 2008 Eighth IEEE International Conference on Data Mining, pp. 413–422. IEEE (2008)
19. Liu, Y., Granier, X.: Online tracking of outdoor lighting variations for augmented reality with moving cameras. IEEE Trans. Visual Comput. Graph. **18**(4), 573–580 (2012)
20. Lu, B.V., Kakuta, T., Kawakami, R., Oishi, T., Ikeuchi, K.: Foreground and shadow occlusion handling for outdoor augmented reality. In: 2010 IEEE International Symposium on Mixed and Augmented Reality, pp. 109–118. IEEE (2010)
21. Ma, W.C., Wang, S., Brubaker, M.A., Fidler, S., Urtasun, R.: Find your way by observing the sun and other semantic cues. In: 2017 IEEE International Conference on Robotics and Automation (ICRA), pp. 6292–6299. IEEE (2017)
22. Madsen, C.B., Lal, B.B.: Outdoor illumination estimation in image sequences for augmented reality. GRAPP **11**, 129–139 (2011)
23. Madsen, C.B., Störring, M., Jensen, T., Andersen, M.S., Christensen, M.F.: Real-time illumination estimation from image sequences. In: Proceedings: 14th Danish Conference on Pattern Recognition and Image Analysis, Copenhagen, Denmark, pp. 1–9 (2005)
24. Pedregosa, F., et al.: Scikit-learn: machine learning in Python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
25. Reda, I., Andreas, A.: Solar position algorithm for solar radiation applications. Sol. Energy **76**(5), 577–589 (2004)

26. Schonberger, J.L., Frahm, J.M.: Structure-from-motion revisited. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4104–4113 (2016)
27. Van Dijk, T., de Croon, G.C.H.E.: How do neural networks see depth in single images? In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2183–2191 (2019)
28. Wei, H., Liu, Y., Xing, G., Zhang, Y., Huang, W.: Simulating shadow interactions for outdoor augmented reality with RGBD data. IEEE Access **7**, 75292–75304 (2019)
29. Whelan, T., Salas-Moreno, R.F., Glocker, B., Davison, A.J., Leutenegger, S.: Elasticfusion: real-time dense slam and light source estimation. Int. J. Robot. Res. **35**(14), 1697–1716 (2016)
30. Woo, S., Park, J., Lee, J.-Y., Kweon, I.S.: CBAM: convolutional block attention module. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11211, pp. 3–19. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01234-2_1
31. Xiao, J., Ehinger, K.A., Oliva, A., Torralba, A.: Recognizing scene viewpoint using panoramic place representation. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 2695–2702. IEEE (2012)
32. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. arXiv preprint arXiv:1611.05431 (2016)
33. Xiong, Y., Chen, H., Wang, J., Zhu, Z., Zhou, Z.: DSNet: deep shadow network for illumination estimation. In: 2021 IEEE Virtual Reality and 3D User Interfaces (VR), pp. 179–187. IEEE (2021)
34. Zhang, J., Sunkavalli, K., Hold-Geoffroy, Y., Hadap, S., Eisenman, J., Lalonde, J.F.: All-weather deep outdoor lighting estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 10158–10166 (2019)
35. Zhang, K., Li, X., Jin, X., Liu, B., Li, X., Sun, H.: Outdoor illumination estimation via all convolutional neural networks. Comput. Electr. Eng. **90**, 106987 (2021)
36. Zhu, Y., Zhang, Y., Li, S., Shi, B.: Spatially-varying outdoor lighting estimation from intrinsics. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12834–12842 (2021)