International Journal of Computer Vision (2023) 131:1302–1306
https://doi.org/10.1007/s11263-022-01747-w

CORRECTION

Correction: Spatio-Temporal Outdoor Lighting Aggregation on Image Sequences Using Transformer Networks

Haebom Lee[1,2] · Christian Homeyer[1,3] · Robert Herzog[1] · Jan Rexilius[4] · Carsten Rother[2]

/ Accepted: 29 December 2022 / Published online: 19 January 2023

Correction: International Journal of Computer Vision https://doi.org/10.1007/s11263-022-01725-2

This erratum aims to correct errors in the sections 1, 3, and 5 of Lee et al. (2022). Some of the texts in these sections were reproduced in non-final form. It resulted in omissions of several major extensions that are made during the revision process. Figures and Tables are not affected.

1. Introduction

- The following sentence in the 3rd paragraph (Fig. 1) should be modified as:

In our work, we go in a similar direction as we robustly estimate the global sun direction and other lighting parameters (Lalonde & Matthews, 2014) by fusing estimates both from the spatial and temporal domains.

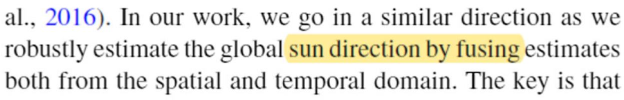

al., 2016). In our work, we go in a similar direction as we robustly estimate the global sun direction by fusing estimates both from the spatial and temporal domain. The key is that

Figure 1 Text to be modified

- The 5th paragraph (Fig. 2) should be modified as:
  … which accounts for individual orientations and field-of-views of the input frames. With this novel pipeline, we eliminate the necessity of intricate hyperparameter tuning required for post-processing. In our experiments in Sect. 4, we replace parts of our estimation pipeline and adapt the architecture of Dosovitskiy et al. (2020) for lighting source regression. To the best of our knowledge, we are the first to use an attention-based model for the task of lighting estimation. Finally, we extend our lighting model. Unlike previous work which predicted only the sun direction, the proposed work estimates parameters of the *Lalonde-Matthews* outdoor illumination model (Lalonde & Matthews, 2014).

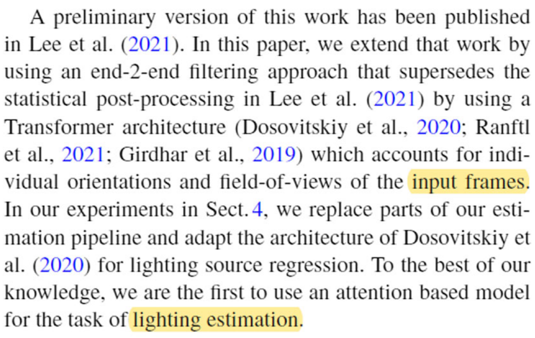

A preliminary version of this work has been published in Lee et al. (2021). In this paper, we extend that work by using an end-2-end filtering approach that supersedes the statistical post-processing in Lee et al. (2021) by using a Transformer architecture (Dosovitskiy et al., 2020; Ranftl et al., 2021; Girdhar et al., 2019) which accounts for individual orientations and field-of-views of the input frames. In our experiments in Sect. 4, we replace parts of our estimation pipeline and adapt the architecture of Dosovitskiy et al. (2020) for lighting source regression. To the best of our knowledge, we are the first to use an attention based model for the task of lighting estimation.

Figure 2 Text to be modified

- The list of contributions in the 6th paragraph (Fig. 3) should be modified as:

  1. Building on top of our preliminary work, we propose a spatio-temporal aggregation for sunlight estimation that is trained end-to-end using a *Transformer* architecture.
  2. A novel handcrafted positional encoding tailored to encode the local and global camera angles for spatio-temporal aggregation.
  3. More realistic lighting estimation using the *Lalonde-Matthews* illumination model (Lalonde & Matthews, 2014).
4. Superior performance compared to the state-of-the-art.

The original article can be found online at https://doi.org/10.1007/s11263-022-01725-2.

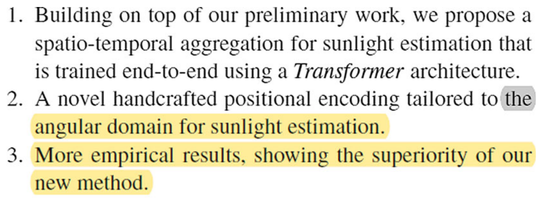
✉ Haebom Lee
haebom.lee@gmail.com

[1] Corporate Research, Robert Bosch GmbH, Hildesheim, Germany

[2] CVL Lab, IWR, Heidelberg University, Heidelberg, Germany

[3] IPA Group, Heidelberg University, Heidelberg, Germany

[4] Bielefeld University of Applied Sciences, Campus Minden, Minden, Germany

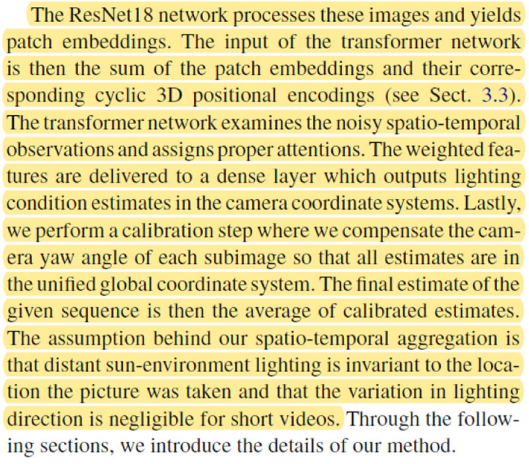


1. Building on top of our preliminary work, we propose a spatio-temporal aggregation for sunlight estimation that is trained end-to-end using a *Transformer* architecture.
2. A novel handcrafted positional encoding tailored to the angular domain for sunlight estimation.
3. More empirical results, showing the superiority of our new method.

Figure 3 Text to be modified

3. Proposed Method

- An additional sentence should be inserted after the last sentence of the 1st paragraph:
  In this way, the samples obtained from each sequence provide different observations for the same global lighting condition. This design is motivated by our empirical results, which showed that lighting can be estimated well from many small parts.
- The 2nd paragraph (Fig. 4) is completely rewritten as:
  All image crops are passed through the backbone network and projected to a sequence of patch embeddings. We then add an orientation-invariant positional encoding and pass the sequence to our transformer network. Through the attention layers, the noisy spatio-temporal observations can be effectively aggregated to a final estimate. Weighted features are delivered to a dense layer that produces the estimated *Lalonde-Matthews* illumination model parameters. The sun direction estimates are formulated in their own camera coordinate systems. We compensate the camera yaw angle of each subimage in order to obtain aligned estimates in a unified global coordinate system. Our final prediction is given as the average of all estimates. Note that the sky parameters of the *Lalonde-Matthews* model do not require the alignment step, as they do not vary with respect to the camera yaw angle. The assumption behind our spatio-temporal aggregation is that distant sun-environment lighting can be considered invariant for small-scale translations (e.g., driving) and that the variation in lighting direction is negligible for short videos. Through the following sections, we introduce the details of our method.



The ResNet18 network processes these images and yields patch embeddings. The input of the transformer network is then the sum of the patch embeddings and their corresponding cyclic 3D positional encodings (see Sect. 3.3). The transformer network examines the noisy spatio-temporal observations and assigns proper attentions. The weighted features are delivered to a dense layer which outputs lighting condition estimates in the camera coordinate systems. Lastly, we perform a calibration step where we compensate the camera yaw angle of each subimage so that all estimates are in the unified global coordinate system. The final estimate of the given sequence is then the average of calibrated estimates. The assumption behind our spatio-temporal aggregation 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 our method.

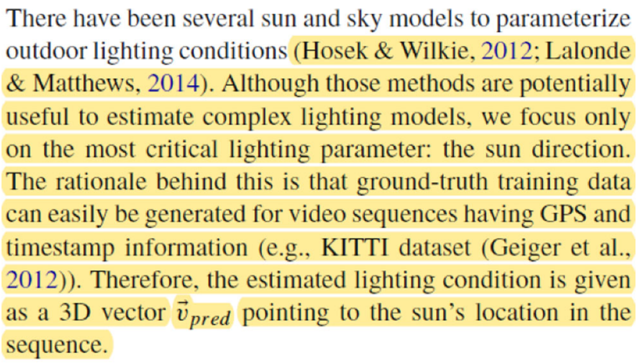
Figure 4 Text to be modified

3.1 Lighting Estimation

- The 1st paragraph (Fig. 5) is completely rewritten as:
  There have been several sun and sky models to parameterize outdoor lighting conditions such as the *Hosek-Wilkie* sky model (Hosek & Wilkie, 2012) or the *Lalonde-Matthews* (Lalonde & Matthews, 2014) outdoor illumination model. In this work, we extend our previous method by predicting the parameters of the *Lalonde-Matthews* model. This hemispherical illumination model (f_{LM}) describes the luminance of outdoor illumination for a light direction l as the sum of sun (f_{sun}) and sky (f_{sky}) components based on 11 parameters:

$$
\begin{aligned}
&f_{LM}(l; q_{LM}) = \boldsymbol{w}_{sun} f_{sun}(l; \beta, \kappa, l_{sun}) + \boldsymbol{w}_{sky} f_{sky}(l; t, l_{sun}),\\
&f_{sun}(l; \beta, \kappa, l_{sun}) = \exp(-\beta \exp(-\kappa/\cos\gamma_l)),\\
&f_{sky}(l; t, l_{sun}) = f_P(\theta_{sun}, \gamma_l, t),\\
&q_{LM} = \left\{\boldsymbol{w}_{sun}, \boldsymbol{w}_{sky}, \beta, \kappa, t, \boldsymbol{l}_{sun}\right\},
\end{aligned}
$$

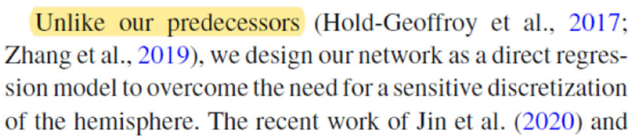
where $\boldsymbol{w}_{sun} \in R^3$ and $\boldsymbol{w}_{sky} \in R^3$ are the mean sun and sky colors, (β, κ) are the sun shape descriptors, t is the sky turbidity, $\boldsymbol{l}_{sun} = [\theta_{sun}, \phi_{sun}]$ is the sun position, γ_l is the angle between the light direction l and the sun position l_{sun}, and f_P is the Preetham sky model (Preetham et al., 1999). For more details, please refer to (Lalonde & Matthews, 2014).



There have been several sun and sky models to parameterize outdoor lighting conditions (Hosek & Wilkie, 2012; Lalonde & Matthews, 2014). Although those methods are potentially useful to estimate complex lighting models, 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 having GPS and timestamp information (e.g., KITTI dataset (Geiger et al., 2012)). Therefore, the estimated lighting condition is given as a 3D vector $\vec{v}_{pred}$ pointing to the sun's location in the sequence.

Figure 5 Text to be modified

- The following sentence should be inserted at the beginning of the 2nd paragraph (Fig. 6):
  Among the parameters, the sun direction may be the most critical component. Unlike our predecessors …



Unlike our predecessors (Hold-Geoffroy et al., 2017; Zhang et al., 2019), 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. (2020) and

Figure 6 An additional sentence should be attached

- The loss functions for the extended sun sky model should be attached at the end of Sect. 3.1 (below Eq. 4) as a new paragraph:

For the remaining parameters, we apply the mean squared error (MSE) to the predicted values and the normalized ground truth values as in Jin et al. (2020):

$$\mathrm{L}_{\mathrm{w_{sun}}} = \frac{1}{3}\|w_{sun}^{pred} - w_{sun}^{gt}\|_2^2$$

$$L_{w_{sky}} = \frac{1}{3}\|w_{sky}^{pred} - w_{sky}^{gt}\|_2^2$$

$$L_{beta} = \|\beta^{pred} - \beta^{gt}\|_2^2$$

$$L_{kappa} = \|\kappa^{pred} - \kappa^{gt}\|_2^2$$

$$L_t = \|t^{pred} - t^{gt}\|_2^2$$

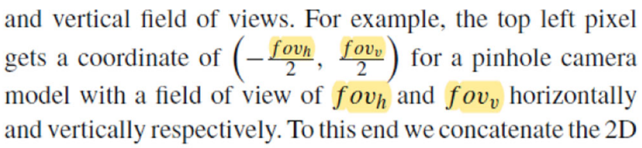
$$L_{param} = \frac{1}{5}\left[L_{w_{sun}} + L_{w_{sky}} + L_{beta} + L_{kappa} + L_t\right]$$

Since the two loss functions L_{sun} and L_{param} have similar magnitudes, we define the final loss function as the sum of them:

$$L_{light} = L_{sun} + L_{param}.$$

3.3 Orientation-Invariant Positional Encoding

- The occurrences of an abbreviation fov (field of view) in the 1st paragraph (Fig. 7) should be substituted with a spherical angle symbol $\sphericalangle$:
  For example, the top left pixel gets a coordinate of $\left(-\frac{\sphericalangle_h}{2}, \frac{\sphericalangle_v}{2}\right)$ for a pinhole camera model with a field of view of $\sphericalangle_h$ and $\sphericalangle_v$ horizontally and vertically, respectively.



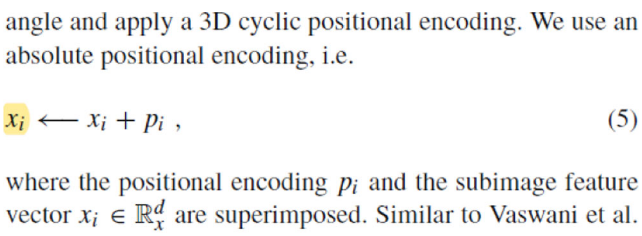
and vertical field of views. For example, the top left pixel gets a coordinate of $\left(-\frac{fov_h}{2}, \frac{fov_v}{2}\right)$ for a pinhole camera model with a field of view of fov_h and fov_v horizontally and vertically respectively. To this end we concatenate the 2D

Figure 7 Text to be modified

- The first occurrence of x_i in the equation 5 (Fig. 8) should be substituted with x_i^{enc}:
  We use an absolute positional encoding, i.e.

$$x_i^{enc} \leftarrow x_i + p_i,$$

where the positional encoding p_i and the subimage feature vector $\mathrm{x_i} \in \mathbb{R}_x^d$ are superimposed.



angle and apply a 3D cyclic positional encoding. We use an absolute positional encoding, i.e.

$$x_i \longleftarrow x_i + p_i \ , \qquad (5)$$

where the positional encoding p_i and the subimage feature vector $x_i \in \mathbb{R}_x^d$ are superimposed. Similar to Vaswani et al.

Figure 8 Text to be modified

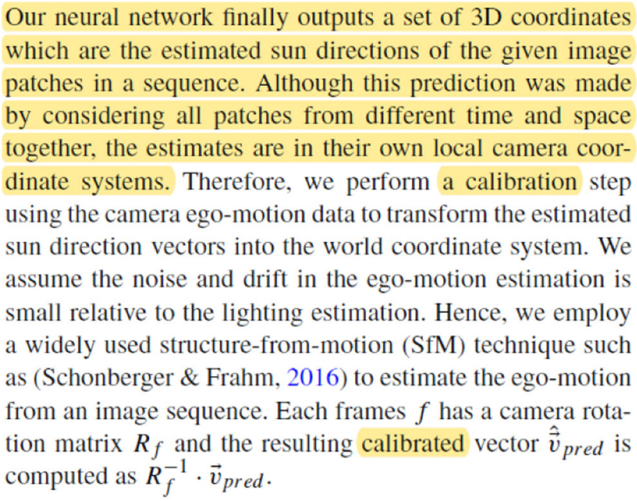
- The following sentence should be inserted after the last sentence:
  The resulting positional encoding of a subimage is the stacked vector of the three cyclic positional encodings. Note that the depth parameter d is carefully determined so that the depth of the stacked vector matches the channel size of the transformer network.

3.4 Calibration

- Occurrences of 'calibration' and 'calibrated' should be substituted with 'alignment' and 'aligned'. This change includes the subsection title.
- The first two sentences are completely rewritten to reflect the changes introduced by an extended sun and sky model.
- A new sentence is inserted at the end of the 1st paragraph.
  The correct text for these three changes is:
  3.4 Alignment

Our neural network outputs the lighting parameters as a 11-dimensional vector for a given sequence of image patches. Although this prediction was made by considering patches from different temporal and spatial locations, the sun direction estimates are in their own local camera coordinate systems. Therefore, we perform an alignment step using the camera ego-motion data 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. Therefore, we employ a widely used structure-from-motion (SfM) technique such as Schonberger & Frahm (2016) to estimate the egomotion of an image sequence.

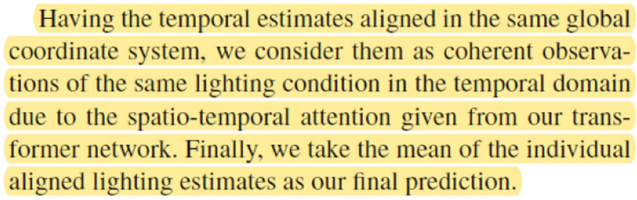
Each frame f has a camera rotation matrix R_f and the resulting aligned vector $\overrightarrow{v}_{pred}$ is computed as $R_f^{-1} \cdot \overrightarrow{v}_{pred}$. Finally, we take the mean of the aligned lighting estimates as our final prediction.



Our neural network finally outputs a set of 3D coordinates which are the estimated sun directions of the given image patches in a sequence. Although this prediction was made by considering all patches from different time and space together, the estimates are in their own local camera coordinate systems. Therefore, we perform a calibration step using the camera ego-motion data 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 widely used structure-from-motion (SfM) technique such as (Schonberger & Frahm, 2016) to estimate the ego-motion from an image sequence. Each frames f has a camera rotation matrix R_f and the resulting calibrated vector $\hat{\vec{v}}_{pred}$ is computed as $R_f^{-1} \cdot \vec{v}_{pred}$.

Figure 9 Text to be modified

- The second paragraph should be removed.



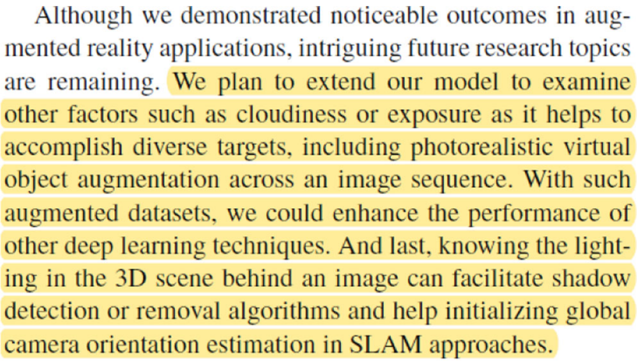
Having the temporal estimates aligned in the same global coordinate system, we consider them as coherent observations of the same lighting condition in the temporal domain due to the spatio-temporal attention given from our transformer network. Finally, we take the mean of the individual aligned lighting estimates as our final prediction.

Figure 10 Text to be deleted

5. Conclusion

- The 2nd paragraph (Fig. 11) is completely rewritten as:

Although we demonstrated visually appealing results in augmented reality applications, intriguing future research topics are remaining open. Intuitively, the performance of the model should scale with the sequence length, as more information is present. We plan to scale both our model and data to examine the limit of attention-based spatio-temporal aggregation for lighting estimation. Another interesting direction would be the integration of our method into reconstruction pipelines, such as SLAM. Knowing the lighting direction and shadow-casting can help initializing camera estimation. Lastly, we want to investigate further into the sampling methods. Instead of picking 8 random frames from an image sequence, we could think of selecting consecutive frames and experiment with the number of frames and the distance from the starting point.



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 across an image sequence. With such augmented datasets, we could enhance the performance of other deep learning techniques. And last, knowing the lighting in the 3D scene behind an image can facilitate shadow detection or removal algorithms and help initializing global camera orientation estimation in SLAM approaches.

Figure 11 Text to be modified

References

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.
2. Hosek, L., & Wilkie, A. (2012). An analytic model for full spectral sky-dome radiance. *ACM Transactions on Graphics (TOG)*.
3. Jin, X., Deng, P., Li, X., Zhang, K., Li, X., Zhou, Q., et al. (2020). Sun-sky model estimation from outdoor images. *Journal of Ambient Intelligence and Humanized Computing*.

4. Lalonde, J.-F., & Matthews, I. (2014). Lighting estimation in outdoor image collections. *IEEE International Conference on 3D Vision*.
5. Preetham, A., Shirley, P., & Smits, B. (1999). A practical analytic model for daylight. In *Proceedings of the 26th annual conference on Computer graphics and interactive techniques*.
6. Schonberger, J., & Frahm, J.-M. (2016). Structure-from-motion revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*.