International Journal of Computer Vision (2023) 131:1060–1072
https://doi.org/10.1007/s11263-022-01725-2

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]

Received: 14 February 2022 / Accepted: 18 November 2022 / Published online: 15 December 2022

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 lighting of the scene in the form of shading and shadows. Recovering the lighting is an inverse rendering problem and as that ill-posed. Recent research based on deep neural networks has shown promising results for estimating light from a single image, but with shortcomings in robustness. We tackle this problem by combining lighting estimates from several image views sampled in the angular and temporal domains of an image sequence. For this task, we introduce a transformer architecture that is trained in an end-2-end fashion without any statistical post-processing as required by previous work. Thereby, we propose a positional encoding that takes into account camera alignment and ego-motion estimation to globally register the individual estimates when computing attention between visual words. We show that our method leads to improved lighting estimation while requiring fewer hyperparameters compared to the state of the art.

Keywords Lighting estimation · Spatio-temporal filtering · Positional encoding · Transformer

1 Introduction

Deep learning models are able to learn strong priors from data for solving highly ill-posed problems like single image reconstruction (Fan et al., 2017). In this manner, they have also been used 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 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 (Balcı & Güdükbay, 2017; Kán & Kaufmann, 2019; Madsen & Lal, 2011; Wei et al., 2019; Zhu et al., 2021). 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 (Van Dijk & de Croon, 2019) that shadows are essential for single-image depth prediction using convolutional neural networks.

Previous methods have focused on estimating explicit sky map textures (Hold-Geoffroy et al., 2019), locating the sun position from single RGB images (Hold-Geoffroy et al., 2017; Jin et al., 2020; Zhang et al., 2019), calculating sun trajectories from longer time-lapse videos (Balcı & Güdükbay, 2017; Liu & Granier, 2012) or in estimating a set of light sources in the context of RGBD-SLAM (Whelan et

Communicated by Michael Möller.

✉ 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 Campus Minden, Bielefeld University of Applied Sciences, Minden, Germany

al., 2016). 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 domain. The key is that we take advantage of known intrinsic calibration and ego-motion of multiple camera images, which all share the same direct sun light that is independent of relative translation. Therefore our method is applicable to both: pure rotational panorama images and images recorded with ego-motion as demonstrated in the results section.

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.). At the same time, not all images in a sequence 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.

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. 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).

We summarize our contributions as follows:

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.

2 Related Work

Estimation of outdoor lighting conditions has been extensively studied due to its importance in computer graphics and computer vision applications (Karsch et al., 2011; Lu et al., 2010). Related techniques can be categorized into two parts, one that analyzes a single image (Hold-Geoffroy et al., 2019, 2017; Jin et al., 2020, 2019; Lalonde et al., 2012; Ma et al., 2017; Zhang et al., 2021) and the other that utilizes a sequence of images (Balcı & Güdükbay, 2017; Lalonde & Matthews, 2014; Liu & Granier, 2012; Madsen & Lal, 2011).

2.1 Single Image

Hold-Geoffroy et al. (2017) proposed a method that estimates outdoor illumination from a single low dynamic range image using a convolutional neural network (Krizhevsky et al., 2012) (CNN). 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 Zhang et al. (2019) where they incorporated a more delicate illumination model (Lalonde & Matthews, 2014). The predicted parameters were evaluated numerically with the ground truth values and rather qualitatively assessed by using the render loss.

Jin et al. (2020) and Zhang et al. (2021) also proposed single image based lighting estimation methods. While their predecessors (Hold-Geoffroy et al., 2017; Zhang et al., 2019) 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. (2021) combined lighting estimation with intrinsic image decomposition. Although they achieved a noticeable result in sun position estimation on synthetic datasets, we could not compare them to ours because their method utilizes intrinsic images which are unavailable for real scene videos.

2.2 Multiple Images

The above lighting estimation methods based on a single image often suffer from insufficient cues to determine a lighting condition, such as when a given image is completely shadowed. Therefore, several attempts were made to increase the accuracy and robustness by taking the temporal

domain into account (Balcı & Güdükbay, 2017; Lalonde & Matthews, 2014; Madsen & Lal, 2011).

For example, in the outdoor illumination estimation method presented by Madsen et al. (2005), the authors estimated the trajectory of the sun and its variable 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 (Kajiya, 1986) optimization problem to determine the continuous change of the lighting parameters. The method introduced in Liu and Granier (2012) extracts a set of features from each image frame and uses it to estimate the relative changes of the lighting parameters in an image sequence. Their method can handle moving cameras and generate time-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 (Reda & Andreas, 2004).

The lighting condition estimation is also crucial in augmented reality where virtual objects are realistic when they are rendered in the background image using the correct lighting conditions. Lu et al. (2010), 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 estimation performance depends solely on the segmentation of the shadow region and the finding of related items. Therefore, the method may be challenging if a shadow-casting object is not visible in the image. Madsen and Lal (2011) utilize a stereo camera to extend (Madsen et al., 2005) further. They estimated sky and sun variations over an image sequence using the sun direction calculated from the GPS coordinates and time stamps. The estimates are then combined with shadow detection algorithms to generate plausible augmented scenes with appropriate shading and shadows.

Recently, several attempts have been made to use auxiliary information to estimate lighting conditions (Kán & Kaufmann, 2019; Xiong et al., 2021). Such information may result in better performance but only with a trade-off in generality. Kán and Kaufmann (2019) proposed a single RGB-D image-based lighting estimation method for augmented reality applications. They used synthetically generated scenes to train a deep neural network that maps the angular coordinates of the main light source in the scene. Outlier removal and temporal smoothing processes were applied to achieve temporal consistency of the method. However, this method was demonstrated only on static-view images. Our method, on the other hand, improves its estimates by aggregating observations from different points of view. 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 two networks: a pre-trained ResNet18 (He et al., 2016) and a transformer network (Vaswani et al., 2017). We first randomly crop several small subimages from a sequence of images. 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. In this way, the samples obtained from each sequence provide different observations for the same global

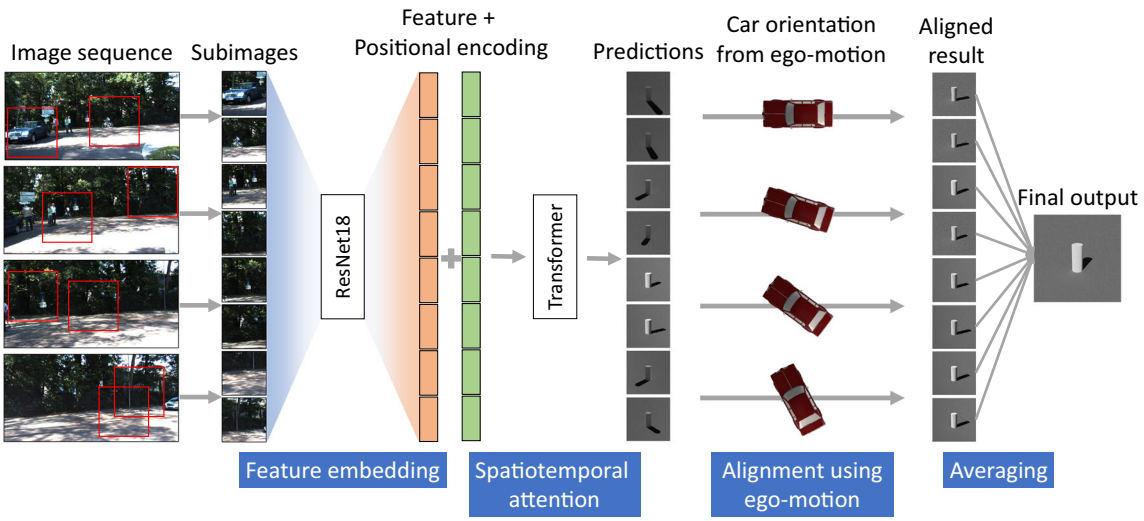


Fig. 1 Spatio-temporal outdoor lighting aggregation on an image sequence: feature vectors are extracted from subimages using a pre-trained ResNet18 network. Using an absolute positional encoding, our transformer network performs spatio-temporal attention. Individual estimates made in each camera coordinate system are calibrated using camera yaw angle data and fused to yield the lighting estimation for the sequence

lighting condition. This design is motivated by our empirical results, which showed that lighting can be estimated well from many small parts.

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.

3.1 Lighting Estimation

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:

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

where $\mathbf{w}_{sun} \in \mathbb{R}^3$ and $\mathbf{w}_{sky} \in \mathbb{R}^3$ are the mean sun and sky colors, (β, κ) are the sun shape descriptors, t is the sky turbidity, $\mathbf{l}_{sun} = [\theta_{sun}, \phi_{sun}]$ 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).

Among the parameters, the sun direction may be the most critical component. 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 Zhang et al. (2021) presented regression networks 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, we compare the estimated sun direction with the ground truth and apply two more conditions to foster the training. The first loss function is defined to minimize the angle between the estimate and the ground truth sun direction $\vec{v}_{gt}$:

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

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:

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

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):

$$
\begin{aligned}
L_{w_{sun}} &= \frac{1}{3}\left\| w_{sun}^{pred} - w_{sun}^{gt} \right\|_2^2\\
L_{w_{sky}} &= \frac{1}{3}\left\| w_{sky}^{pred} - w_{sky}^{gt} \right\|_2^2\\
L_{beta} &= \left\| \beta^{pred} - \beta^{gt} \right\|_2^2\\
L_{kappa} &= \left\| \kappa^{pred} - \kappa^{gt} \right\|_2^2\\
L_t &= \left\| t^{pred} - t^{gt} \right\|_2^2\\
L_{param} &= \frac{1}{5}\left[L_{w_{sun}} + L_{w_{sky}} + L_{beta} + L_{kappa} + L_t \right]
\end{aligned} \tag{5}
$$

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

$$L_{light} = L_{sun} + L_{param}. \tag{6}$$

3.2 Attention Based Aggregation

In order to extract robust estimates from noisy observations, the aggregation process described in Lee et al. (2021) relies heavily on statistical filtering utilizing an outlier removal combined with the meanshift algorithm. However, this approach requires manual hyper-parameter tuning with handcrafted selection criteria. We extend this work by replacing the aggregation step with a purely end-to-end attention driven pipeline. The overview of our approach is illustrated in Fig. 1.

We take inspiration from Dosovitskiy et al. (2020) for our network design and adopt their hybrid architecture for our task. This includes self attention using multi-head attention layers (Vaswani et al., 2017) and preprocessing images with a pretrained convolutional neural network. Given a temporal sequence of k images, we first select n spatially random crops for each frame as done in our previous work (Lee et al., 2021). On each crop, we apply a ResNet18 (He et al., 2016) encoder to extract feature embeddings. Each embedded patch is fed as input to our transformer module for aggregation. The virtue of the transformer network is that it can associate observations from different space and time given a proper positional encoding. Since all images patches share the same sun light and we assume we know their relative orientation due to the ego-motion estimation the Transformers attention mechanism inherently learns to filter the noisy patch-wise predictions. However, we need to provide the relative orientation of the patches in order to make the light estimation invariant to camera orientation, which we achieve via the *positional encoding*.

3.3 Orientation-Invariant Positional Encoding

Solely relying on image features enables only to estimate the lighting in the local camera frame. However, we need to fuse the estimates in a global reference frame in order relate different subimages. Since we assume sun-lighting, only the directional component of a recorded camera image is relevant to calibrate different frames. We inject this camera orientation in the image features via a positional encoding. However, we only encode the yaw angle of the camera rotations (the rotation around the ground-plane surface-normal) since pitch and roll angles are naturally captured in the image features of outdoor images (e.g, horizon). Further, we also encode the 2D position of the subimages cropped from the source frame independent of the intrinsic camera projection, i.e., in terms of viewing angles ϕ in the corresponding horizontal and vertical field of views. 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. To this end we concatenate the 2D angular image coordinate and the (temporal) camera rotation angle and apply a 3D cyclic positional encoding. We use an absolute positional encoding, i.e.

$$x_i^{enc} \longleftarrow x_i + p_i \ , \tag{7}$$

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. (2017) we use a fixed encoding of sine and cosine functions with different frequencies.

Since our positional encoding scheme encodes angles, it has to fulfill the following two conditions: (1) *periodicity*—the transition from the encoding of 359° to the encoding of 0° should be as smooth as the transition from 0° to 1° and (2) *uniqueness*—each angle should have a unique encoding. We present our cyclic positional encoding, satisfying those conditions, by using nested trigonometic functions as below:

$$\begin{aligned} PE\ (\phi,\ 2i) &= \sin\left(\sin\left(\phi\right) \cdot \alpha / 10000^{2i/d}\right) \\ PE\ (\phi,\ 2i+1) &= \sin\left(\cos\left(\phi\right) \cdot \alpha / 10000^{2i/d}\right), \end{aligned} \tag{8}$$

where $i \in [0,\ \frac{d}{2})$ and d denotes the depth of the positional encoding. Note that α is an empirically determined parameter, which controls the width of the nonzero area of the encoding. The periodicity comes from the nested trigonometric function while uniqueness is established by interlacing the two functions. Figure 2 shows the positional encoding generated by the above function.

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 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 location, 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. 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 aligned vector $\hat{\vec{v}}_{pred}$ is computed as $R_f^{-1} \cdot \vec{v}_{pred}$. Finally, we take the mean of the aligned lighting estimates as our final prediction.

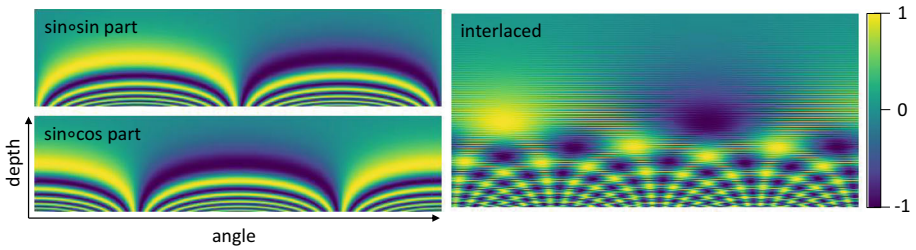


Fig. 2 Cyclic positional encoding for angle $\phi \in [0,\ 2\pi]$. The periodicity of our encoding scheme is clearly visible on the left side images while their interlaced result on the right side shows its uniqueness for each angle

Table 1 Number of data in our datasets

Dataset	Training		Validation		Test	
	Sequences	Images	Sequences	Images	Sequences	Images
SUN360	10000	160000	1000	16000	1000	16000
KITTI	4208	33889	427	3508	432	3457

4 Experiments

4.1 Datasets

We choose two datasets for evaluation: KITTI (Geiger et al., 2012) and SUN360 (Xiao et al., 2012). KITTI is a popular dataset for autonomous driving. It consists of multiple driving sequences with rectified images and has additional annotations for determining the ground-truth sun directions (Reda & Andreas, 2004). This makes it an ideal candidate to test our method on everyday driving scenes. For our experiments we create a random *train-val-test* split composed of 47-5-5 driving scenes. This results in 33 889, 3508, and 3457 images, respectively. Note that this *scene* is different from the *sequence* we give to the network. Since we generate a *sequence* by randomly selecting eight frames from the same *scene* during the training and inference, there are 4208, 427, and 432 sequences for the *train-val-test* split, respectively. (see Table 1). For the sampling in the spatial domain, four subimages are randomly cropped from each frame image while allowing overlapping. Our pipeline estimates the global sun direction from this spatio-temporal sequence of 32 images. Since KITTI does not provide ground truth Lalonde-Matthews lighting model parameters, we omit the loss for other lighting parameters (L_{param}). Therefore, the loss function becomes $L_{light} = L_{sun}$.

The SUN360 dataset is another common dataset considered for outdoor lighting estimation methods because 1) it provides diverse environments and 2) there is a labeling of the parameters of the Lalonde-Matthews lighting model (Zhang et al., 2019). Several previous methods used it in its original panorama form or as subimages by generating synthetic perspective images (Hold-Geoffroy et al., 2017). We followed the latter approach, which has also been used in our preliminary work (Lee et al., 2021) where we examined the performance improvement arising from spatial aggregation.

In this paper, we propose to build an artificial image sequence from a panorama so that we can examine and compare our method's performance with previous works. Specifically, we simulate a camera motion without translation by generating a set of synthetic perspective images with a fixed field of view and randomized camera yaw and pitch angles. By doing so, we can perform the spatio-temporal aggregation on the SUN360 dataset in the same manner as on KITTI. We start with dividing 12 000 panorama images into the training, validation, and test sets with a 10:1:1 ratio. From each panorama, a sequence of 16 perspective images with random yaw angles is generated while allowing overlapping. We want to have the data from both datasets as similar as possible. Therefore, we match the horizontal and vertical field of views and set the numbers of random frames and subframes to 8 and 4 respectively. Since there are 16 frames for each panorama, a sequence of 8 frames has C_8^{16} different combinations, resulting in great diversity. Note that we also introduce small random offsets on the camera elevation with respect to the horizon in $[-10°,\ 10°]$. The generated images are resized to 1220×370 to match the size of the KITTI images. In this way, we produced 160 000, 16 000, and 16 000 images from 10 000, 1000, and 1000 panoramas for training, validation, and test sets, respectively. The exact numbers of panoramas and images are presented in Table 1, and Fig. 3 illustrates examples from the two datasets.

4.2 Implementation Details

As illustrated in Fig. 4, our lighting estimation model consists of a ResNet18 network and a transformer network, followed by dense layers converting a feature vector of dimension 512 to the estimates for the 3D sun direction and other lighting parameters (only applicable to SUN360). It accepts 32 RGB images of size 224×224 cropped from 8 frames and outputs the lighting estimate through the alignment and averaging process. We borrow the core structure of the transformer from

Fig. 3 Examples of the two datasets (Geiger et al., 2012; Xiao et al., 2012). From the original image (*top*), we generate random subimages (*bottom*)

Dosovitskiy et al. (2020) and carefully determine the number of layers, number of heads, hidden size, and MLP size as 4, 4, 512, and 1024, respectively, under extensive experiments. The dropout rate was 0.2.

We train our model and test its performance on the SUN360 and KITTI datasets separately (see Table 1). In detail, we empirically trained our lighting estimation network for 118 and 131 epochs for the SUN360 and KITTI datasets using early stopping. The training was initiated with the AdamW optimizer (Loshchilov & Hutter, 2017) using a learning rate of 1×10^{-5} and the batch size was 8. It took 61.1 and 34.3 h on a single Nvidia RTX 3090 GPU. Prediction on a single sequence of 32 images takes 90 ms. Our spatio-temporal aggregation model is examined on 1000 unobserved SUN360 sequences and 432 KITTI sequences.

4.3 Results

4.3.1 Sun Direction

We evaluate the angular errors of the spatio-temporally aggregated sun direction estimates on the SUN360 test sequences. Since other single image-based lighting estimation methods (Hold-Geoffroy et al., 2017; Jin et al., 2020; Zhang et al., 2021) are not capable of conducting spatio-temporal aggregation, the median of the estimates over each sequence is utilized. On top of that, we compare our method with the spatio-temporal aggregation pipeline proposed in Lee et al. (2021). The hyperparameters required for our previous method are determined in the same way as described in Lee et al. (2021).

Figure 5 illustrates the cumulative angular errors of the five methods trained and tested on the SUN360 dataset. We present the outcomes of three single image based approaches along with the results of two spatio-temporal aggregation methods. Our spatio-temporal attention method shows a noticeable margin compared to the state-of-the-art.

We also performed a similar comparison on the KITTI dataset (see Fig. 6). On this dataset, however, we compare our method only with (Lee et al., 2021) due to the lack of ground truth information such as exposure and turbidity which are required for other previous works. Although the dataset provides the ground truth ego-motion required for the alignment step, we calculated it using (Schonberger & Frahm, 2016) to generalize our approach. The mean angular error of the estimated camera rotation using the default parameters was 1.01° over the five test scenes. Using the proposed spatio-temporal attention method, the mean angular error over the 432 test sequences recorded 7.96°, which is marginally better than 9.62° of Lee et al. (2021).

We plotted the individual sun direction estimates and their aggregation results using our methods and (Lee et al., 2021) in Fig. 7. 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 lighter color dots. The single image estimation of Lee et al. (2021) was performed individually and resulted in independent noisy estimates which were aggregated by statistical post-processing. Unlike them, our estimates are jointly predicted and therefore tend to cluster tightly around their mean rendering any statistical post-processing redundant. The mean standard deviation of sun direction estimations also demonstrates our model's capability for coherent estimation (see Figs. 5 and 6). Compared to other methods, we recorded 2 to 6 times lower mean standard deviation. This behavior comes from the spatio-temporal attention from our transformer network. We contend that the network tries to output a set of predictions that can explain the lighting condition of the given sequence, rather than predicting each subimage's lighting condition individually. Furthermore, this characteristic supports our decision to average all estimates to obtain the final estimate of the sequence.

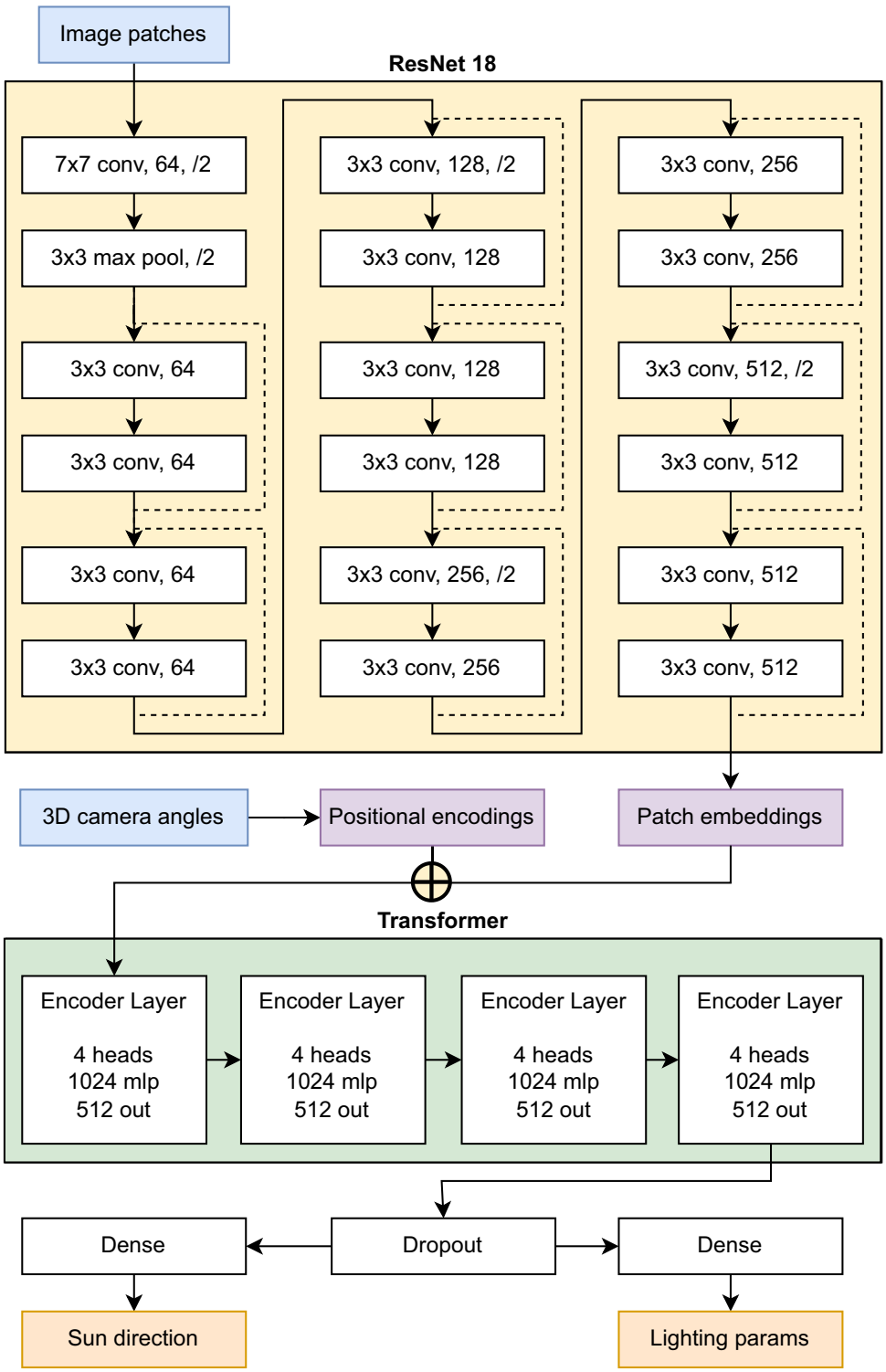
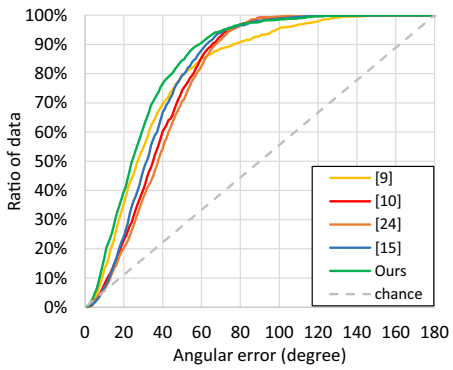
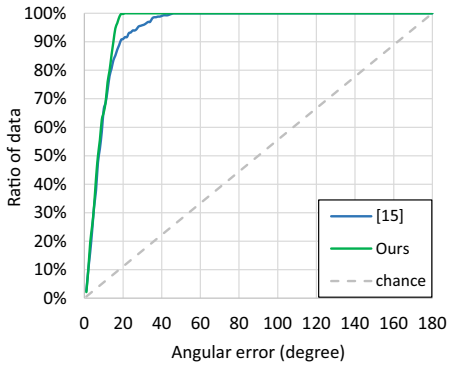


Fig. 4 The proposed lighting estimation model. The features of the input image patches are extracted through the ResNet18 (He et al., 2016) network. We generate orientation-invariant positional encodings from the given 3D camera angles and add them (denoted as $\oplus$) to the patch embeddings. Our transformer network then aggregates the observations and outputs the estimated sun direction and lighting parameters of the sequence. Note that the right-side dense layer is omitted for the KITTI dataset



	Median	Mean	Min	Max	SD
[9]	27.00	35.39	1.27	161.03	0.3325
[10]	35.01	37.36	0.84	**118.10**	0.1895
[24]	37.75	39.12	**0.21**	126.65	0.1915
[15]	31.66	35.20	0.33	137.57	0.1297
Ours	**24.12**	**29.41**	0.78	143.47	**0.0569**

Fig. 5 The cumulative angular error and the statistics of the sun direction estimates on the SUN360 test set. (Lee et al., 2021) and *Ours* are showing the spatiotemporal aggregation results. For a fair comparison, angular errors of other methods are measured upon the median of the estimates made on single images. The proposed method outperforms other methods with a noticeable margin. Values of the best method/setting for each column are given in bold



	Median	Mean	Min	Max	SD
[15]	7.42	9.62	**0.23**	45.93	0.0838
Ours	**7.04**	**7.96**	0.55	**20.42**	**0.0142**

Fig. 6 The cumulative angular error and the statistics on the KITTI test set. Our method performs slightly better than (Lee et al., 2021) while recording a noticeable small maximum angular error of 20.42°. Values of the best method/setting for each column are given in bold

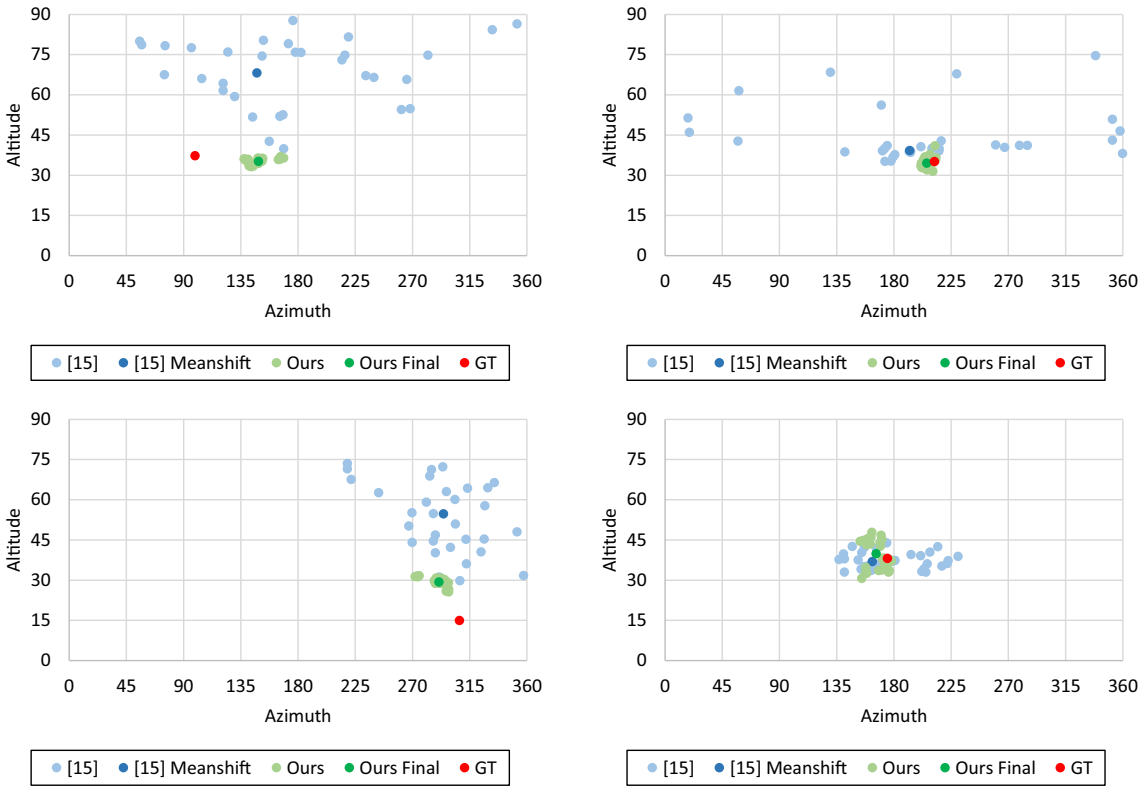


Fig. 7 Scatter plots representing sun direction estimates of individual subimages and the spatiotemporal aggregation result. Each plot corresponds to an image sequence of 8 frames in (*left*) the SUN360 and (*right*) the KITTI test sets. The spatio-temporal aggregation proposed in Lee et al. (2021) finds the highest point density among the inliers treating the estimates as independent sample. On the contrary, individual estimates of our method form a tight group due to the spatio-temporal attention

Table 2 RMSE of the estimated parameters on the SUN360 test set

	$\mathbf{w}_{sun}$	$\mathbf{w}_{sky}$	κ	β	t
Hold-Geoffroy et al. (2017)	–	–	–	–	1.0869
Jin et al. (2020)	0.3680	0.1083	0.1817	9.7960	1.1994
Ours	**0.2810**	**0.0833**	**0.1201**	**6.9778**	**0.9510**

Values of the best method/setting for each column are given in bold

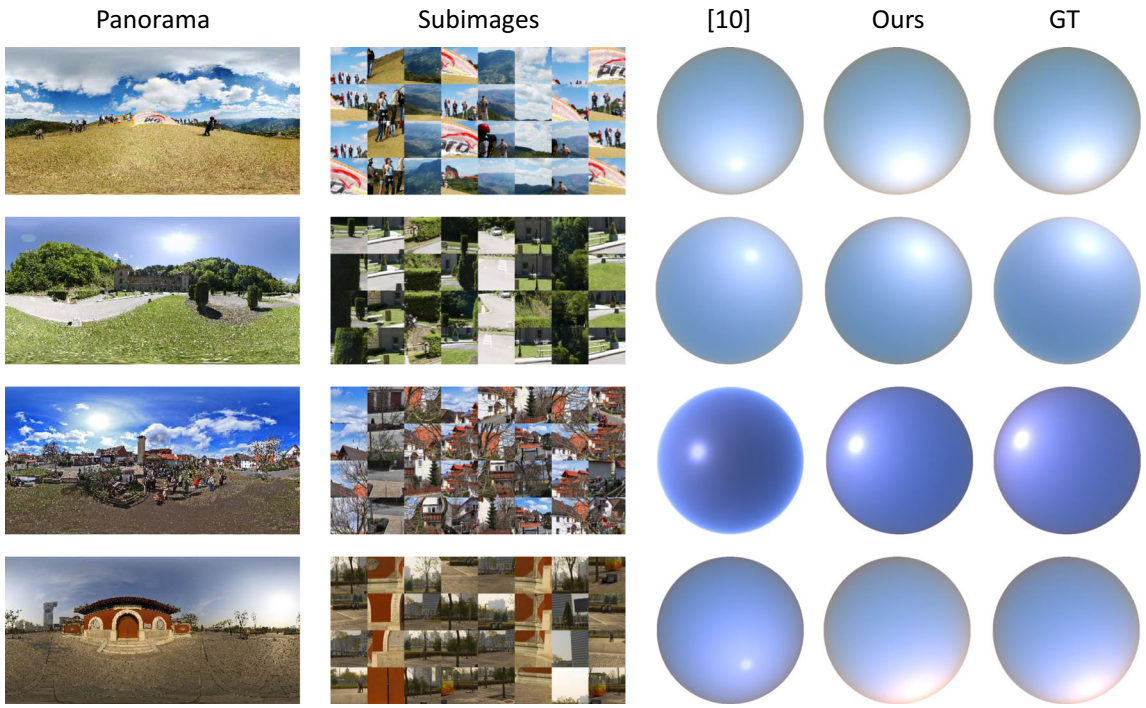


Fig. 8 Qualitative comparison on the estimated parameters of the *Lalonde-Matthews* model. Our methods aggregates information obtained from the subimages of a synthetic sequence and provides plausible outcomes on various lighting conditions

4.3.2 Other Lighting Parameters

As described earlier, the remaining *Lalonde-Matthews* model's parameters are only estimated for the SUN360 dataset. We present the root mean squared errors of Hold-Geoffroy et al. (2017), Jin et al. (2020), and ours in Table. 2. Note that (Hold-Geoffroy et al., 2017) only delivers the RMSE for turbidity, because it is based on a different lighting model. Our method demonstrated outstanding performance for all five items. We also provide a qualitative evaluation on the full *Lalonde-Matthews* model in Fig. 8. Each hemispherical texture is generated using the estimated/ground truth parameters.

The stability of our model is better understood with a virtual object augmentation application, as shown in Fig. 9. Note that other lighting parameters, such as the sun's intensity, are manually determined and equally applied for the single image estimation method and (Lee et al., 2021). 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 almost entirely removed and the augmented object's appearance is almost identical to the ground truth after applying the spatio-temporally aggregated lighting condition based on the *Lalonde-Matthews* model.

4.4 Ablation Study

We perform a series of ablations for our chosen losses, positional encoding and the number of patches for our model. Ablations are done on the SUN360 test set and we compare angular error statistics.

4.4.1 Loss Function

Table 3 shows the angular error statistics for different loss term combinations. The L_{cosine} metric was set as the default loss function as it dominantly drives the training. Best performance can be achieved by using all loss terms together.

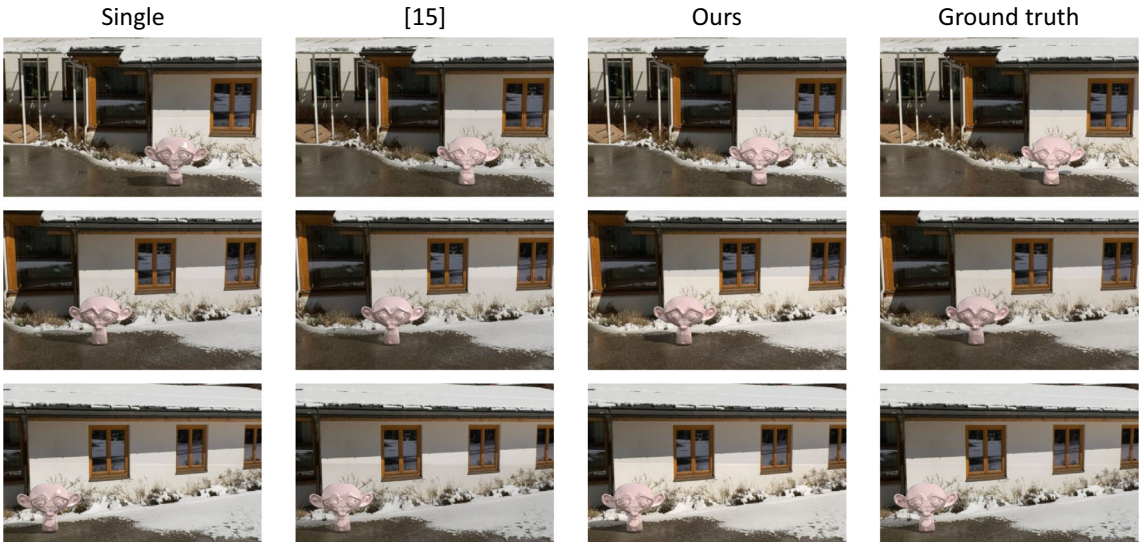


Fig. 9 Virtual Augmentations: Fluctuations in the shadow of the augmented object are strongly visible when the sun direction is estimated individually. Our spatio-temporal method (Lee et al., 2021) achieves more stable results. The proposed learned aggregation results in even better quality, almost indistinguishable from the ground truth

Table 3 Ablation study with loss functions on the SUN360 test set

L_{cosine}	L_{norm}	L_{hemi}	Median	Mean	Min	Max
✓			26.20	31.47	1.24	157.98
✓	✓		25.00	30.59	**0.29**	157.96
✓		✓	25.04	30.94	0.51	157.65
✓	✓	✓	**24.12**	**29.41**	0.78	**143.47**

Values of the best method/setting for each column are given in bold

Table 4 Ablation study with positional encoding schemes on the SUN360 test set

	Median	Mean	Min	Max
None	35.56	37.99	1.30	157.11
Standard	27.42	32.06	**0.55**	165.64
Ours	**24.12**	**29.41**	0.78	**143.47**

Values of the best method/setting for each column are given in bold

4.4.2 Positional Encoding

We investigate the benefit of our newly proposed orientation-invariant positional encoding by comparing it to the standard sinusodial encoding introduced in Vaswani et al. (2017). The results in Table. 4 show, that our task-specific encoding gives greater performance over the standard one or using none at all.

4.4.3 Patch Sequence

In these experiments, we ablate the number and choice of patches given to the aggregation transformer. By changing the number of frames and number of spatial patches per image, we compare different temporal-spatial patch variations. The results in Table 5 show that there is a sweet spot for the length of the temporal sequence and the number of patches per image. We achieve the best performance by choosing a sequence of 8 images and 4 patches per image, resulting in a sequence length of 32. Increasing the sequence length seems to hurt the model performance at a certain point. We believe that this could be due to the limited model capacity and plan to experiment with larger networks in the future.

Table 5 Ablation study with hyperparameters on the SUN360 test set

Frames	Subimages	Median	Mean	Min	Max
4	4	25.83	31.31	**0.66**	155.42
8	4	**24.12**	**29.41**	0.78	**143.47**
12	4	24.62	30.33	0.97	151.44
16	4	25.91	31.02	1.11	160.40
8	2	24.87	30.82	**0.58**	160.38
8	4	**24.12**	**29.41**	0.78	**143.47**
8	6	24.53	30.60	0.80	173.33
8	8	25.55	31.29	0.91	152.70

Values of the best method/setting for each column are given in bold

5 Conclusion

In this paper, we proposed a holistic sequence-wise lighting estimation method based on spatio-temporal attention using transformers. Our method achieved state-of-the-art perfor-

mance on outdoor lighting estimation for a given image sequence. Without loss of generality we utilized 360° panoramas and wide view images in our work, but the method can also be applied to any image providing enough details. Moreover, our spatio-temporal aggregation could also be generalized to other globally shared image information under given computational budgets.

Although we demonstrated noticeable outcomes 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.

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Balcı, H., & Güdükbay, U. (2017). Sun position estimation and tracking for virtual object placement in time-lapse videos. *Signal, Image and Video Processing, 11*(5), 817–824.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S., et al. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Fan, H., Su, H., & Guibas, L. J. (2017). A point set generation network for 3d object reconstruction from a single image. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 605–613).

Geiger, A., Lenz, P., & Urtasun, R. (2012). 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.

Girdhar, R., Carreira, J., Doersch, C., & Zisserman, A. (2019). Video action transformer network. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (pp. 244–253).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 770–778).

Hold-Geoffroy, Y., Athawale, A., & Lalonde, J.-F. (2019). Deep sky modeling for single image outdoor lighting estimation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 6927–6935).

Hold-Geoffroy, Y., Sunkavalli, K., Hadap, S., Gambaretto, E., & Lalonde, J.-F. (2017). Deep outdoor illumination estimation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 7312–7321).

Hosek, L., & Wilkie, A. (2012). An analytic model for full spectral sky-dome radiance. *ACM Transactions on Graphics (TOG), 31*(4), 1–9.

Jin, X., Deng, P., Li, X., Zhang, K., Li, X., Zhou, Q., Xie, S., & Fang, X. (2020). Sun-sky model estimation from outdoor images. *Journal of Ambient Intelligence and Humanized Computing*, (pp. 1–12).

Jin, X., Sun, X., Zhang, X., Sun, H., Xu, R., Zhou, X., et al. (2019). Sun orientation estimation from a single image using short-cuts in dcnn. *Optics & Laser Technology, 110,* 191–195.

Kajiya, J. T. (1986). The rendering equation. In: *Proceedings of the 13th annual conference on computer graphics and interactive techniques*, (pp. 143–150).

Kán, P., & Kaufmann, H. (2019). Deeplight: Light source estimation for augmented reality using deep learning. *The Visual Computer, 35*(6–8), 873–883.

Karsch, K., Hedau, V., Forsyth, D., & Hoiem, D. (2011). Rendering synthetic objects into legacy photographs. *ACM Transactions on Graphics (TOG), 30*(6), 1–12.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In: *Advances in Neural Information Processing Systems*, (pp. 1097–1105).

Lalonde, J.-F., & Matthews, I. (2014). Lighting estimation in outdoor image collections. In: *2014 2nd international conference on 3D vision*, vol. 1, (pp. 131–138). IEEE.

Lalonde, J.-F., Efros, A. A., & Narasimhan, S. G. (2012). Estimating the natural illumination conditions from a single outdoor image. *International Journal of Computer Vision, 98*(2), 123–145.

Lee, H., Herzog, R., Rexilius, J., & Rother, C. (2021). Spatiotemporal outdoor lighting aggregation on image sequences. In: *DAGM German conference on pattern recognition*, (pp. 343–357). Springer.

Liu, Y., & Granier, X. (2012). Online tracking of outdoor lighting variations for augmented reality with moving cameras. *IEEE Transactions on visualization and computer graphics, 18*(4), 573–580.

Loshchilov, I., & Hutter, F. (2017). Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101.

Lu, B. V., Kakuta, T., Kawakami, R., Oishi, T., & Ikeuchi, K. (2010). Foreground and shadow occlusion handling for outdoor augmented reality. In: *2010 IEEE International symposium on mixed and augmented reality*, (pp. 109–118). IEEE.

Ma, W.-C., Wang, S., Brubaker, M.A., Fidler, S., & Urtasun, R. (2017). 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.

Madsen, C. B., Störring, M., Jensen, T., Andersen, M. S., & Christensen, M. F. (2005). Real-time illumination estimation from image sequences. In: *Proceedings: 14th Danish conference on pattern recognition and image analysis*, Copenhagen, Denmark, (pp. 1–9).

Madsen, C. B., & Lal, B. B. (2011). Outdoor illumination estimation in image sequences for augmented reality. *GRAPP, 11,* 129–39.

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 (Vol. 99, pp. 91–100).

Ranftl, R., Bochkovskiy, A., & Koltun, V. (2021). Vision transformers for dense prediction. In: *Proceedings of the IEEE/CVF international conference on computer vision*, (pp. 12179–12188).

Reda, I., & Andreas, A. (2004). Solar position algorithm for solar radiation applications. *Solar Energy, 76*(5), 577–589.

Schonberger, J. L., & Frahm, J.-M. (2016). Structure-from-motion revisited. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 4104–4113).

Van Dijk, T., & de Croon, G. C. H. E. (2019). How do neural networks see depth in single images? In: *Proceedings of the IEEE international conference on computer vision*, (pp. 2183–2191).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In: *Advances in Neural Information Processing Systems*, (pp. 5998–6008).

Wei, H., Liu, Y., Xing, G., Zhang, Y., & Huang, W. (2019). Simulating shadow interactions for outdoor augmented reality with rgbd data. *IEEE Access, 7,* 75292–75304.

Whelan, T., Salas-Moreno, R. F., Glocker, B., Davison, A. J., & Leutenegger, S. (2016). Elasticfusion: Real-time dense slam and light source estimation. *The International Journal of Robotics Research, 35*(14), 1697-1716.

Xiao, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2012). Recognizing scene viewpoint using panoramic place representation. In: *2012 IEEE conference on computer vision and pattern recognition*, (pp. 2695–2702). IEEE.

Xiong, Y., Chen, H., Wang, J., Zhu, Z., & Zhou, Z. (2021). Dsnet: Deep shadow network for illumination estimation. In: *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, (pp. 179–187). IEEE.

Zhang, J., Sunkavalli, K., Hold-Geoffroy, Y., Hadap, S., Eisenman, J., & Lalonde, J.-F. (2019). All-weather deep outdoor lighting estimation. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, (pp. 10158–10166).

Zhang, K., Li, X., Jin, X., Liu, B., Li, X., & Sun, H. (2021). Outdoor illumination estimation via all convolutional neural networks. *Computers & Electrical Engineering, 90,* 106987.

Zhu, Y., Zhang, Y., Li, S., & Shi, B. (2021). Spatially-varying outdoor lighting estimation from intrinsics. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, (pp. 12834–12842).