High-Dynamic-Range Lighting Estimation From Face Portraits

Alejandro Sztrajman
University College London
London, UK
a.sztrajman@ucl.ac.uk

Alexandros Neophytou
Microsoft
Reading, UK
alexandros.neophytou@microsoft.com

Tim Weyrich
University College London
London, UK
t.weyrich@cs.ucl.ac.uk

Eric Sommerlade
Microsoft
Reading, UK
eric.sommerlade@microsoft.com

Abstract

We present a CNN-based method for outdoor high-dynamic-range (HDR) environment map prediction from low-dynamic-range (LDR) portrait images. Our method relies on two different CNN architectures, one for light encoding and another for face-to-light prediction. Outdoor lighting is characterised by an extremely high dynamic range, and thus our encoding splits the environment map data between low and high-intensity components, and encodes them using tailored representations. The combination of both network architectures constitutes an end-to-end method for accurate HDR light prediction from faces at real-time rates, inaccessible for previous methods which focused on low dynamic range lighting or relied on non-linear optimisation schemes. We train our networks using both real and synthetic images, we compare our light encoding with other methods for light representation, and we analyse our results for light prediction on real images. We show that our predicted HDR environment maps can be used as accurate illumination sources for scene renderings, with potential applications in 3D object insertion for augmented reality.

1. Introduction

Light prediction from images enables many applications, such as 3D object insertion for augmented reality [3], material appearance estimation [17] and reflectance transfer [16]. In the case of face images, an accurate and fast estimation of the illumination can be leveraged for real-time realistic 3D object insertion and editing of scene attributes [11], with applications on portrait images and video conferencing systems. While lighting estimation from faces has already been studied [11, 20], previous works focused on predicting low-frequency and low dynamic range lighting, or were based on non-linear optimisation [1], which is time-consuming at inference and prone to local minima. Outdoor lighting is characterised by an extremely high dynamic range, due to the gap in intensity between the sun and the rest of the environment map, and thus its prediction requires an accurate encoding of a wide range of intensity levels.

We present a method for face lighting estimation which takes as input a single portrait image and predicts the corresponding outdoor high-dynamic-range (HDR) environment map. Our method is based on the combination of two deep convolutional network architectures, one for light encoding and the other for face-to-light prediction. While previous methods have attempted to infer the illumination by using faces as light probes [1], the prediction of the high-intensity pixels is usually inaccurate, which makes the resulting HDR maps inappropriate for changes of exposure level and its use as realistic lighting sources in renderers. Our proposed method relies on using different representations for the low and high-intensity pixels of an HDR map, thus improving the light prediction and producing environment maps that can be used to insert rendered photo-realistic 3D assets in the scene.

The key contributions of our work are as follows:

- We define a new compact encoding for outdoor HDR environment maps, which focuses on generating an accurate representation of the high-intensity illumination, while also preserving a realistic appearance of the low intensity.
- We implement two CNN architectures, one for light encoding and another one for light prediction from portrait images. Together these form an end-to-end network architecture able to predict HDR environment maps from low-dynamic-range (LDR) portrait images

at much faster rates than previous methods [1], with potential applications for real-time light prediction and 3D object insertion.

In Section 5.1, we compare our light encoding with other methods for light representation, both in terms of environment map reproduction and as sources of illumination for 3D scenes. In Section 5.2, we show the results of our method for light prediction, applied to both synthetic and real portrait images. Furthermore, we show that our predicted HDR environment maps can be used as visually plausible illumination sources for scene rendering.

2. Related work

2.1. Outdoor light prediction

Light prediction from a single outdoor image was first proposed by Lalonde *et al*. [12]. More recent approaches have leveraged the use of convolutional neural networks to predict outdoor HDR environment maps from LDR images. Zhang *et al*. [26] proposed a deep autoencoder framework to regress HDR environment maps from LDR panoramic images, effectively learning an inverse tone mapping for outdoor scenes. Hold-Geoffroy *et al*. [6, 7] and Zhang *et al*. [27] developed neural methods to predict outdoor HDR environment maps from limited field-of-view portions of LDR panoramic images. Georgoulis et al. [4] used specular objects of uniform reflectance and known geometry as probes to estimate the lighting and material properties. Similarly, LeGendre *et al*. [14] infer plausible HDR illumination from indoor and outdoor LDR images from a mobile phone. Training data is collected from videos that include visible spheres of different reflectances on screen, and then used to implement an image-based loss by differential relighting.

Closest to our work, Calian *et al*. [1] use human faces as probes to predict the illumination of the scene. While they employ a similar autoencoder architecture to generate the environment map encodings, their method relies on a non-linear optimisation which is time-consuming at inference and prone to local minima. In contrast, our method performs the light prediction in an end-to-end network framework, resulting in much faster inference.

Yi *et al*. [25] leverage a large dataset of faces for unsupervised training of a network for highlight extraction. A second architecture maps the highlights to a parametric environment map, enabling HDR light estimation from indoor and outdoor portrait images. Sun *et al*. [23] use a light stage setup to capture 18 subjects and generate relighted portrait images using a large database of environment maps. They use this data to train a convolutional architecture for fast light prediction and relighting of portrait images.

Zhou *et al*. [29] use a similar architecture for light prediction and portrait relighting, but they leverage a large dataset of faces with lower quality light estimation, and include a GAN loss to correct inaccuracies in the dataset. Our own architecture for light prediction is inspired by the encoder of the Hourglass architecture used by Zhou *et al*. [29]. As in their implementation, we predict lighting from the luminance of a portrait image; however, we have replaced their spherical harmonics light representation by a custom-encoding.

2.2. Illumination encoding

Light estimation from images is an ill-posed inverse problem, due to the ambiguity in the decomposition of light and reflectance contributions [24]. This problem is usually addressed by using a constrained model to encode the illumination, such as Spherical Harmonics [15, 11, 21, 29], as originally proposed by Ramamoorthi and Hanrahan [18, 19]. In the case of outdoor illumination, multiple analytical models have been proposed to represent the sky hemisphere. Zhang *et al*. [27] perform a comparison of the Hošek-Wilkie model for clear skies [8, 9] and the Lalonde-Matthews sky model [13], as priors for a CNN architecture that predicts outdoor HDR environment maps from LDR images in all-weather conditions. Calian *et al*. [1] use the Lalonde-Matthews sky model [13] to predict illumination from faces, and compare it to a data-driven representation generated by a convolutional autoencoder, shown to produce more accurate predictions. Our method lies at the intersection between both representations, using a parametric model for the sun and an autoencoder-based encoding for the rest of the illumination. This split allows us to correctly reconstruct the high-intensity peak of the sun, while also providing flexibility in the modelling of the rest of the illumination, including the sky and the ground. The reconstruction of the lower hemisphere is addressed by Calian *et al*. (2018) by modelling the ground with a uniform albedo, and it is not discussed by Zhang *et al*. (2019), as they focus on the upper hemisphere.

3. Method

In this section we present the general method with a detailed description of the architectures of the networks used for light encoding and for the prediction of the incident lighting at a face.

3.1. HDR environment map Encoding

In order to encode our outdoor HDR environments maps we split the HDR light data L in two parts L_{low} and L_{high}, corresponding to the low ($0 \leq L \leq 1$) and high intensity values:

$$L_{low} = \texttt{clip}(L, 0, 1) \tag{1}$$

$$L_{high} = L - L_{low} \tag{2}$$

The high-intensity part L_{high} usually contains a small number of very bright pixels corresponding to the sun, as shown in the top row of Figure 1. In the bottom row we show renderings of a synthetic face with the corresponding environment maps from the top row. While the few bright pixels from L_{high} *seem* negligible in an LDR environment map, they contribute a significant part of the illumination of the face. Moreover, its accurate representation is critical to reconstruct the full dynamic range of the scene, capturing the correct lighting at different exposure levels [2].

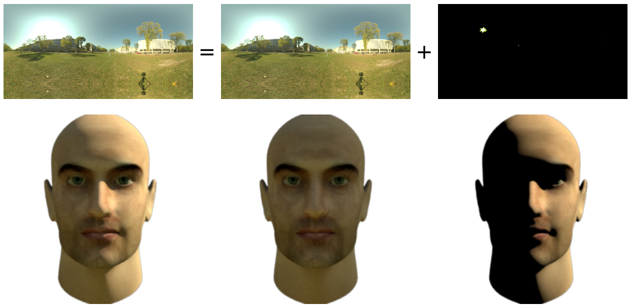

Figure 1. *Top*: splitting of outdoor HDR environment map: $L = L_{\text{low}} + L_{\text{high}}$. *Bottom*: Synthetic face rendered with each corresponding environment map (L, L_{low}, L_{high}).

To encode L_{low} we use a deep convolutional autoencoder architecture with a 16-values embedding layer as shown in Figure 2, trained with outdoor HDR environment maps collected from the *Laval Outdoor HDR Dataset* [6].

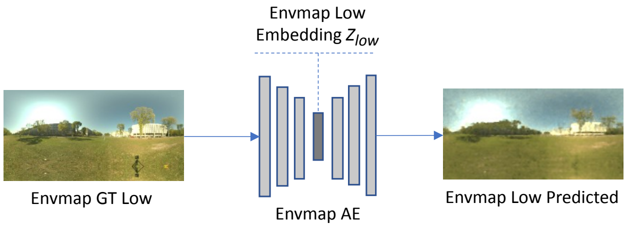

Figure 2. EnvNet: autoencoder of L_{low}.

To encode L_{high} we fit three 2D Gaussian models (one per colour channel) using Levenberg-Marquardt. This allows us to model a single high-intensity light source. In cases with additional sources, such as environments with strong reflections from windows or water surfaces, our model is expected to gradually degrade, toward modelling these sources as LDR content. In each Gaussian model we fit the two coordinates of the centre and the amplitude, and we set the covariance to a fixed value. While this value can also be successfully fitted during the optimisation, we found its predicted value to have very little spread, suggesting that most of the time the fixed-size sun disk overpowers surrounding scattering.

The optimisation is initialised using the maximum RGB values in the environment map as starting points for the Gaussian centres and amplitudes, leading to a quick convergence. The resulting representation for L_{high} contains **5**-parameters corresponding to the two coordinates of the Gaussian centre and the amplitudes for each colour channel. Putting both parts together our full representation for environment maps is a 21-value embedding (16 for L_{low} + 5 for L_{high}), which we will use in our face-to-light architecture.

3.2. Face to Light

In Figure 3, we detail the architectures of the environment map autoencoder and the network for light prediction from faces.

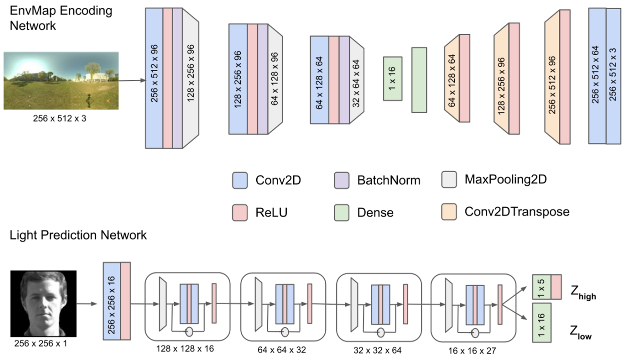

Figure 3. Architectures of the networks for environment map encoding and prediction from faces.

The light prediction network takes as input the luminance of a background-less portrait image (256 × 256 × 1) and outputs a 21-parameters embedding corresponding to the predicted illumination encoding Z_{env}. This consists on the concatenation of the encodings of the low (Z_{low}) and high (Z_{high}) parts of the predicted outdoor HDR environment map. Although using only the luminance has the potential of missing information from the scene, leading to ambiguities in the prediction, especially in regards of L_{low}, this provided more stable predictions than using the full RGB input.

3.3. Training

Training of the environment map prediction network is performed with real face-environment map pairs from the *Laval face+lighting HDR dataset* and with synthetic data generated by combining scanned faces from the *ICT 3D Relightable Facial Expression Database* [22] (ICT-3DRFE) with environment maps from the *Laval Face+Lighting dataset* [1] (see Section 3.4). The encodings used to represent the environment maps are obtained by applying the procedure from Section 3.1. In the case of real portrait images, we perform a background segmentation and removal by using Mask R-CNN [5]. The full pipeline is displayed in Figure 4.

The training loss applied to the environment map encoding is MSE, however different scaling factors are used for different segments of z_{env}, corresponding to the environment map low

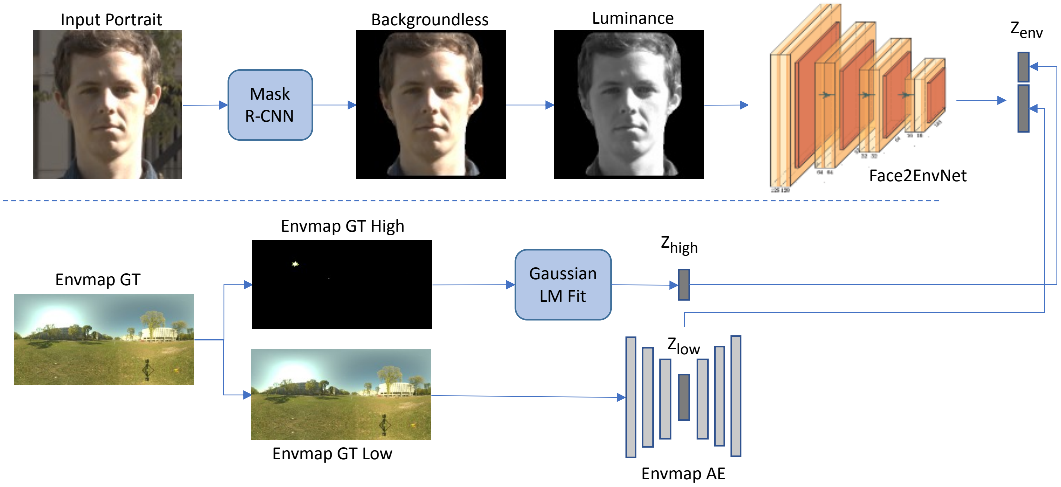

Figure 4. Full processing pipeline of portrait and environment map images, for the training of *Face2EnvNet*.

part z_{low} and two different parts of the Gaussian parameters z_{high}: the amplitudes and the centre position. For the EnvNet autoencoder, the training loss applied to the HDR image data is MAE.

The prediction of the sun's azimuth angle requires special attention in the face-to-light network. A straightforward MSE loss in training leads to inaccurate predictions in cases where the sun is located near the edge of the environment map image. This can be solved by modifying the loss to enforce the periodicity of the azimuth parameter:

$$\text{LOSS}(\theta, \theta^*) = \text{MSE}(\cos\theta, \cos\theta^*) + \text{MSE}(\sin\theta, \sin\theta^*) \quad (3)$$

The environment map prediction network has 5.4×10^5 parameters and is trained for 350 epochs with the ADAM optimiser, using a mini-batch size of 15 and an initial learning rate of 0.0002. The training of both networks takes approximately an hour on a GeForce RTX 2080 Ti GPU, and inference takes approximately 0.5 ms.

3.4. Datasets

The dataset used to train the *EnvNet* autoencoder is the *Laval Outdoor HDR dataset* [6], which consists of 200 HDR environment maps of different outdoor locations and weather conditions. The dataset was augmented by generating random azimuth rotations of the original environment maps, as shown in Figure 5, generating a total of 2, 200 samples. In addition, the environment maps were further multiplied by a random scaling factor between 0.1 and 1 to compensate for the lack of dark environment maps in the dataset (e.g. cases where the sun is hidden or very low in the horizon).

The training of *Face2EnvNet* involved the use of real and synthetic face-environment map pairs.

Real samples came from the *Laval Face+Lighting HDR dataset*, which consists of portrait images of 9 subjects illuminated under 25 different lighting conditions, and their corresponding HDR environment maps, for a total of 137 face/lighting pairs, as shown in Figure 6. This dataset was augmented by horizontal flipping of both face and environment map images.

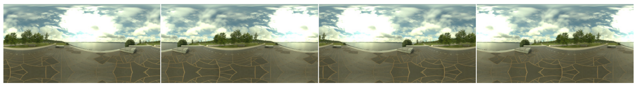

Figure 5. Environment map augmentation: random rotations of the azimuth angle.

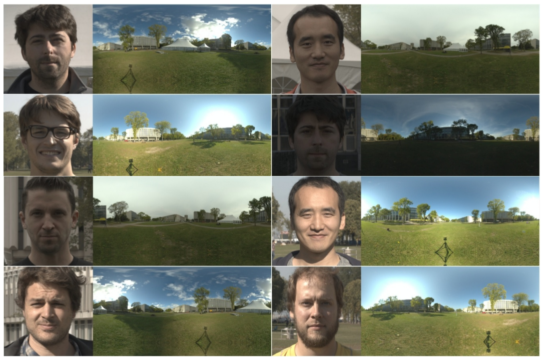

Figure 6. Overview of *Laval Face+Lighting dataset* [1], which includes real faces captured outdoors under different illumination conditions, and their corresponding HDR environment maps.

Synthetic data was generated by combining 30 faces from the *ICT 3D Relightable Facial Expression Database* [22] (ICT-3DRFE), with randomly chosen environment maps from the *Laval Face+Lighting HDR dataset* augmented with azimuth rotations to a total of 500 environment maps. Additionally, small variations of scene parameters were randomly introduced, such as face rotations in the three axes and distance to the camera. Although this was done to introduce robustness in the training, the light prediction network was only tested with front-facing real portraits. Images were rendered using the Mitsuba Renderer [10] for a total of 1000 synthetic face/environment map pairs, such as shown in Figure 7.

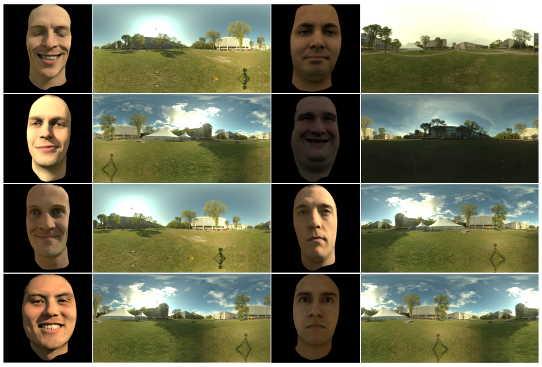

Figure 7. Overview of synthetic dataset generated from combining scanned faces from the *ICT 3D Relightable Facial Expression Database* [22] (ICT-3DRFE) with environment maps from the *Laval Face+Lighting dataset* [1].

4. Evaluation metrics

Similarly to Calian *et al*. [1], we evaluate the quality of reconstruction of HDR environment maps through two different sets of metrics: lighting-based metrics (MAE-dω and RMSE-dω) directly quantify the image difference between ground truth and predicted environment maps, weighed by the solid angle subtended by each pixel in the full sphere or the upper hemisphere (Sky MAE–dω). Shading-based metrics on the other hand, measure the quality of environment map reconstruction indirectly, by computing image losses (MAE, RMSE, SSIM, LPIPS) over a set of scenes rendered using the ground truth and predicted environment maps as sources of illumination. This set of metrics emphasises the use of predicted environment maps as lightmaps for rendering applications, such as 3D object insertion.

5. Results

5.1. EnvNet Encoding

In Figure 8 we evaluate the performance of our method (AE+GAUSS) for light encoding (described in Section 3.1). Training was done with the *Laval Outdoor HDR dataset* [6], augmented with random azimuth rotations as described in Section 3.4. In Figure 8 we show the encoding of 6 previously unseen HDR environment maps, and compare our encoding (21 parameters) with the results produced by fitting the ground truth environment maps to 2nd (SH9) and 4th (SH25) order spherical harmonics, with 27 and 75 parameters correspondingly. We also show the results of encoding the entire HDR environment map using an autoencoder (AE), without separation between L_{low} and L_{high}, and removing the Gaussian fit. Each 5×3 block in Figure 8 shows different encodings fitted to the same environment map with, and below them the results of using each encoding to illuminate two scenes, one with a purely diffuse 3D dragon, and one with a very specular one. Table 1 summarises the results of our encodings comparison, with statistics taken over 41 previously unseen HDR environment maps from the *Laval Outdoor HDR dataset* [6].

The HDR environment maps reconstructed from the spherical-harmonics representations present an abstract similarity to the ground truth, while our encoding recovers a discernible appearance which allows us to identify general characteristics of the scene, such as the colour of the floor, the weather conditions and the location of the sun. This is reflected in the lighting-based metrics from Table 1, where SH errors more than double our encoding's loss. In the case of shading errors, the differences between methods are smaller, although our method still manages to achieve smaller errors using a much smaller number of encoding parameters. A qualitative analysis of the dragon renderings from Figure 8 shows that our method is consistently better at preserving shadows, associated with the directional light L_{high} from the sun, and often also better captures the overall illumination colour of the scene, predominantly linked to L_{low}.

For a fixed exposure, a pure AE encoding can provide a seemingly more accurate reconstruction than the AE+GAUSS method, such as shown in Figure 9. However, the full dynamic range of the sun's intensity is lost in the AE encoding, leading to renderings that are too dark and have lost their dominant directional light. The separation of the environment map light into L_{low} and L_{high}, and its disjoint encoding (AE+GAUSS) explained in Section 3.1, produces a more accurate HDR representation of the illumination source, leading to renderings that preserve shadows and better match the brightness of the scene.

5.2. Light Prediction

In Figures 10 and 11 we show the results of our full pipeline for light prediction from faces (Figure 4) applied to synthetic and real portrait images. On the left we show the input portrait image, in the center we show ground truth and predicted environment maps, and on the right we present renderings of a synthetic face using GT and the predicted environment maps as illumination source. For real portrait images (Figure 11) we further compare our results with predictions from Calian *et al*.'s method [1], kindly provided by the authors. To account for potential differences in radiometric scale between the HDR datasets, we renormalised Calian *et al*.'s environment maps by a global multiplier, chosen to minimise the average RMSE with respect to the ground truth *Laval Face+Lighting dataset* [1], thus granting their results to most favourable comparison to our

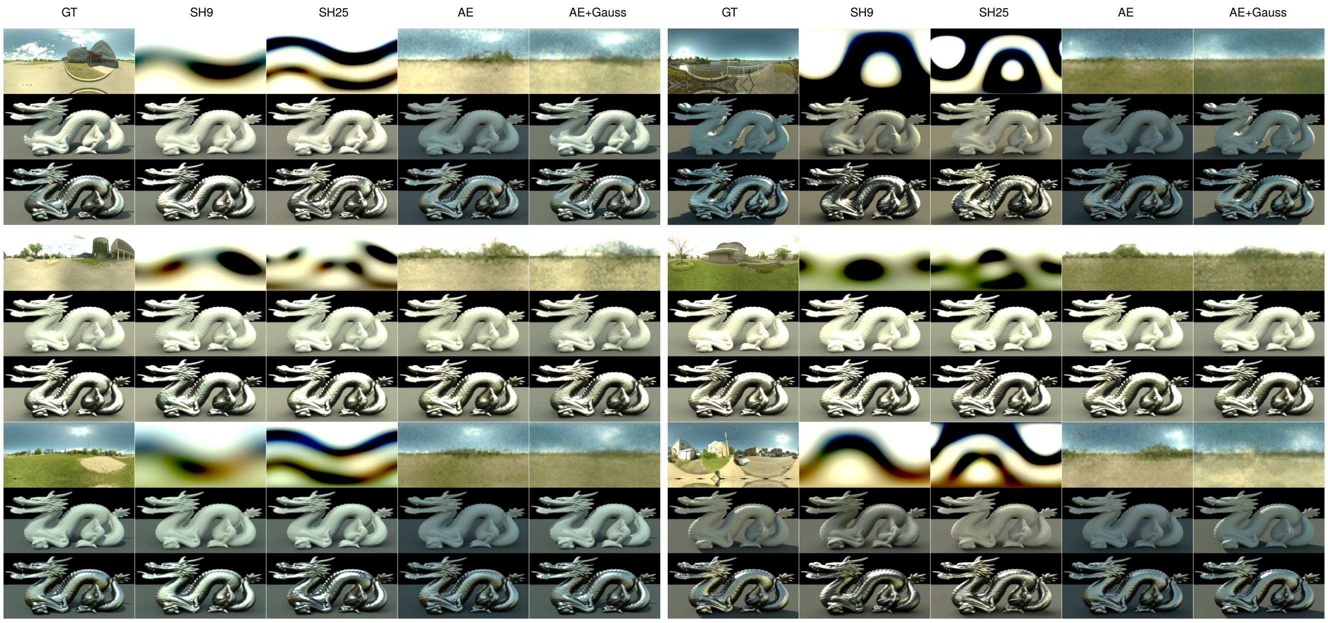

Figure 8. Comparison of outdoor environment maps fitted with our *AE+Gauss* encoding, with spherical harmonics of degrees 2 (SH9) and 4 (SH25), and with an autoencoder without split between L_{low} and L_{high}. Each 5×3 block shows the fitting of one environment map with the different encodings, and the results of using each fitted environment map as illumination to render scenes with diffuse and specular dragons. *Left to right*: GT, SH9, SH25, AE, AE+Gauss. *Top to bottom*: environment map, rendering of diffuse asset, rendering of specular asset.

	MAE	RMSE	Lighting MAE–dω	Lighting RMSE–dω	# Parameters
SH9	0.030 ± 0.024	0.031 ± 0.025	0.20 ± 0.09	0.21 ± 0.09	18
SH25	0.025 ± 0.024	0.026 ± 0.022	0.21 ± 0.11	0.21 ± 0.11	75
AE	0.032 ± 0.024	0.033 ± 0.024	0.12 ± 0.04	0.12 ± 0.04	**16**
AE+Gauss	**0.022 ± 0.015**	**0.023 ± 0.015**	**0.08 ± 0.03**	**0.08 ± 0.03**	21

Table 1. Quantitative comparison of light encoding methods. Lighting-based and shading-based metrics (see Section 4) taken over 41 previously unseen HDR environment maps from the *Laval Outdoor HDR dataset* [6].

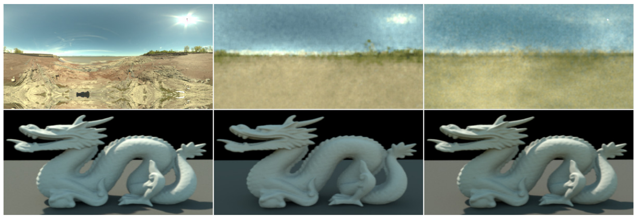

Figure 9. GT Environment map (*left*) reconstructed from AE (*center*) and ae+gauss (*right*) encodings. *Bottom*: renderings of diffuse dragon scene with corresponding environment map. The AE reconstruction looks accurate at a fixed exposure level, but the full dynamic range of the sun is lost in the encoding, producing renderings that are too dark and have no directional light.

method. A quantitative comparison, using both lighting-based and shading-based metrics, is summarised in Tables 2 and 3. The full set of results can be found in the *additional material*, together with the pre-trained networks for light encoding and prediction.

In general, our network's predictions show a good agreement with the ground-truth environment maps, both for real and synthetic portrait images. There is a good visual match between environment map images, signalling an adequate prediction of the L_{low} encoding parameters. A comparison of the rendered faces on the right also shows a generally good prediction of the sun position and intensity, encoded by the L_{high} parameters. Additionally, there is good agreement in the prediction of the sun position in both parts of the encoding: L_{low} and L_{high} consistently describe the sun in close locations.

In terms of lighting-based metrics, shown in Table 2, Calian *et al.*'s method provides a closer prediction of the environment maps, with the exception of the sun's altitude which is consistently placed very low, close to the horizon line (see Figure 11). In contrast, the comparison of shading-based metrics from Table 3 shows that our method outperforms Calian *et al.*'s both in point-wise and perceptually-based metrics. This is evidenced in the better reconstruction of

	MAE–dω	RMSE–dω	Sky MAE–dω	Sun Altitude (rad)	Sun Azimuth (rad)
Our method → Real	0.14 ± 0.07	0.15 ± 0.08	0.27 ± 0.14	**0.12** ± **0.09**	0.48 ± 0.66
Calian *et al.* (scaled) → Real	**0.12** ± **0.06**	**0.13** ± **0.06**	**0.22** ± **0.11**	0.41 ± 0.20	**0.12** ± **0.11**

Table 2. Quantitative results for light prediction from real faces. Comparison of lighting-based metrics (see Section 4) of our method with Calian *et al.* [1]. Calian *et al.*'s predicted environment maps have been scaled to match the average RMSE of the ground truth *Laval Face+Lighting HDR dataset* [1].

	MAE	RMSE	SSIM	LPIPS v0.1 [28]
Our method → Real	**0.015** ± **0.010**	**0.017** ± **0.011**	**0.95** ± **0.02**	**0.029** ± **0.018**
Calian *et al.* (scaled) → Real	0.027 ± 0.016	0.031 ± 0.019	0.89 ± 0.03	0.086 ± 0.014

Table 3. Quantitative results for light prediction from real faces. Comparison of shading-based metrics (see Section 4) of our method with Calian *et al.* [1], taken over a rendered 3D face, as seen on the right side of Figure 11. Calian *et al.*'s predicted environment maps have been scaled to match the average RMSE of the ground truth *Laval Face+Lighting HDR dataset* [1].

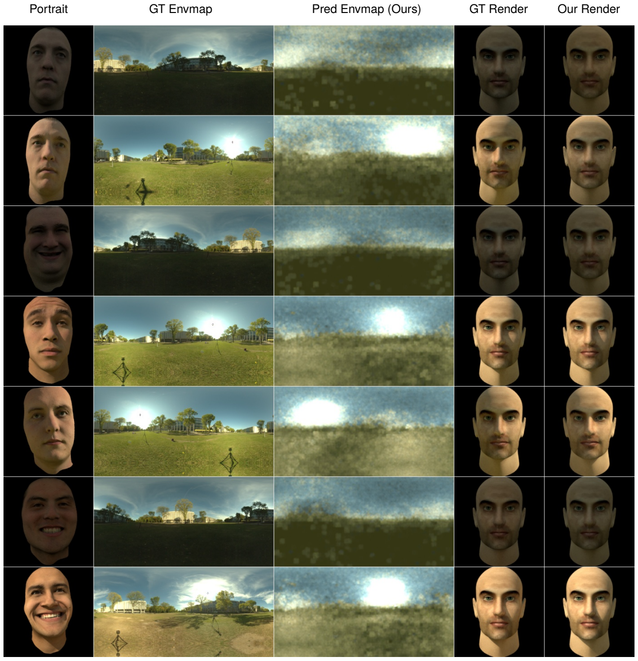

Figure 10. Results on 7 synthetic faces from the *ICT 3D Relightable Facial Expression Database* [22] (ICT-3DRFE) illuminated with environment maps from the *Laval Face+Lighting dataset* [1]. *Left to right*: input portrait image, GT environment map, predicted environment map, synthetic face rendered with GT environment map and with predicted environment map.

shadows and overall illumination colour in the renderings from Figure 11.

Although the *light encoding network* was trained with a wide variety of environment maps, the *light prediction network* is biased towards the appearance of a university campus, as a consequence of our limited training data. Given the robustness of the system, however, we believe that a more varied training dataset would mitigate such bias.

Noticeably, our predictions are good for different weathers, including overcast skies and dark environment maps, where the sun is hidden or very low in the sky. This can be observed in Figure 10, rows 1 and 3, and Figure 11, row 7. The *Laval Outdoor HDR Dataset* [6], used to train our environment map autoencoder, does not contain samples with such low-dynamic-range, and hence the intensity augmentation of the dataset described in Section 3.4 is crucial for the correct encoding and prediction in these cases.

In cases where the light is frontal (sun in the middle of the environment map image), small errors in the prediction of the sun position can lead to noticeable changes in the direction of the shadows on the face. This can be observed in the first row of Figure 11.

Finally, Figure 11 row 4 shows a failure case in the estimation of directional light. Although the L_{low} reconstruction is close to the ground truth, the network predicts the existence of two bright sources of illumination in the sky, and places the L_{high} peak near the wrong one, effectively rotating the directional light source. Similar examples can be observed in the supplemental material, along with the details of the Gaussian reconstruction of L_{high}.

6. Conclusions

We presented a method for outdoor HDR environment map prediction from portrait images, relying on two CNN architectures, for light encoding and light prediction. Our proposed representation for light leverages the extreme dynamic range in outdoor scenes to generate a compact encoding. By combining both networks we generate an end-to-end pipeline for accurate HDR light estimation from faces, with an inference time of 0.5 ms, suitable for real-time applications. We compared our light encoding with other methods for light representation, showing that it is able to produce more realistic and compact representations of HDR lighting,

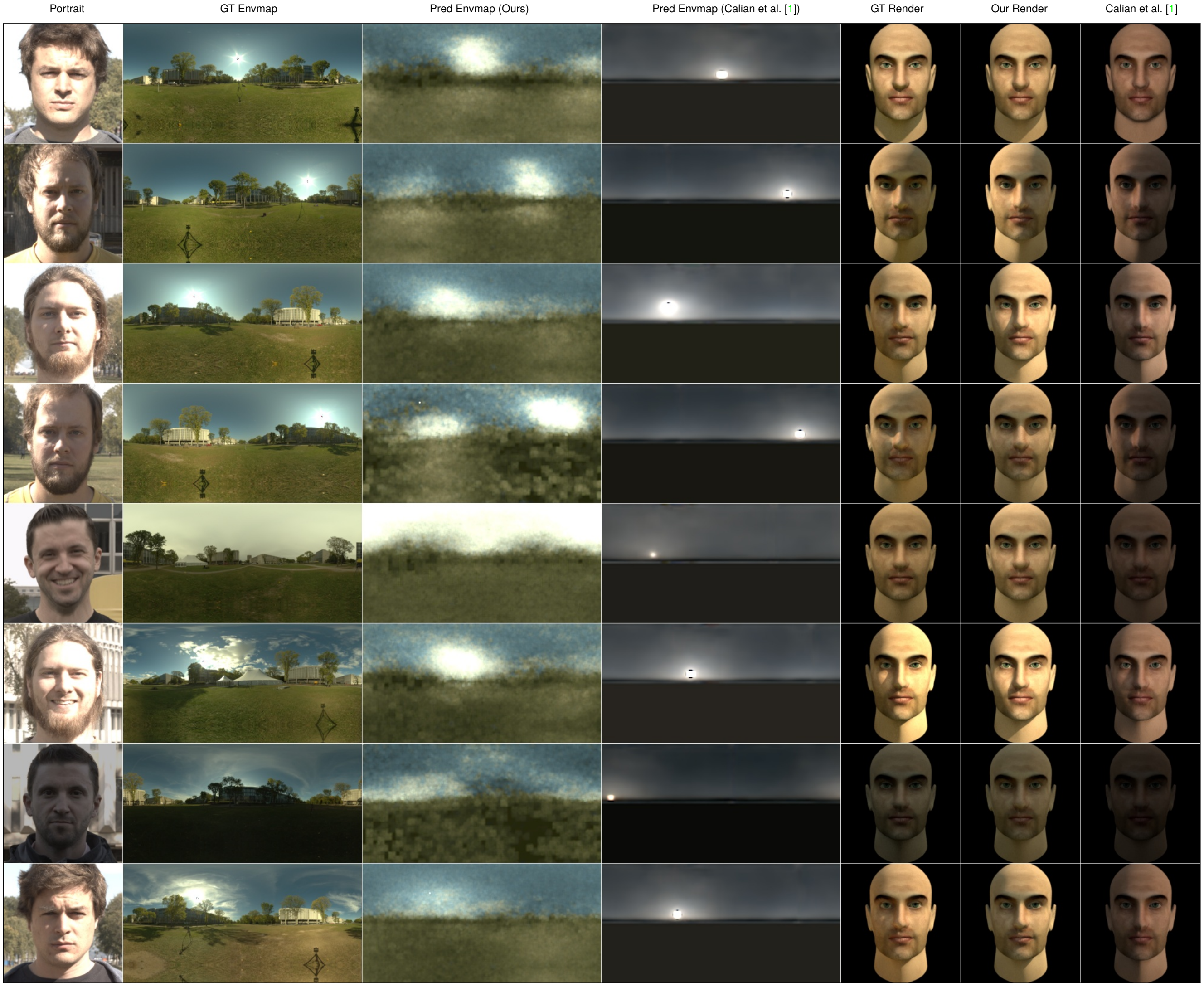

Figure 11. Results on 8 real faces from the *Laval Face+Lighting HDR dataset* and comparison with predictions from Calian *et al*. [1]. *Left to right*: input portrait image, GT environment map, predicted environment map with our method and Calian *et al*.'s, synthetic face rendered with GT environment map and with predicted environment maps.

while also accurately preserving the information necessary for rendering applications. We analysed the estimation of light from real and synthetic portrait images, showing that the low-intensity prediction preserves a realistic look for the environment map, while the high-intensity prediction accurately predicts the intensity and position of the sun. As a result, this better preserves light reflections and hard shadows, conveying a realistic illumination consistent with the original portrait image, required for applications such as object insertion and face relighting.

Acknowledgements

This project has received funding from the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 642841.

References

[1] D. A. Calian, J.-F. Lalonde, P. Gotardo, T. Simon, I. Matthews, and K. Mitchell. From Faces to Outdoor Light Probes. *Computer Graphics Forum*, 37, 2018. 1, 2, 3, 4, 5, 7, 8

[2] P. E. Debevec and J. Malik. Recovering high dynamic range radiance maps from photographs. In *Proceedings of SIGGRAPH 1997*, page 369–378, 1997. 3

[3] M.-A. Gardner, K. Sunkavalli, E. Yumer, X. Shen, E. Gambaretto, C. Gagné, and J.-F. Lalonde. Learning to predict indoor illumination from a single image. *ACM Trans. Graph.*, 36(6), Nov. 2017. 1

[4] S. Georgoulis, K. Rematas, T. Ritschel, E. Gavves, M. Fritz, L. Van Gool, and T. Tuytelaars. Reflectance and natural

illumination from single-material specular objects using deep learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(8):1932–1947, 2018. 2

[5] K. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2980–2988, 2017. 3

[6] Y. Hold-Geoffroy, A. Athawale, and J.-F. Lalonde. Deep sky modeling for single image outdoor lighting estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6920–6928, 06 2019. 2, 3, 4, 5, 6, 7

[7] Y. Hold-Geoffroy, K. Sunkavalli, S. Hadap, E. Gambaretto, and J. Lalonde. Deep outdoor illumination estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2373–2382, 2017. 2

[8] L. Hosek and A. Wilkie. An analytic model for full spectral sky-dome radiance. *ACM Trans. Graph.*, 31:95:1–95:9, 2012. 2

[9] L. Hosek and A. Wilkie. Adding a solar-radiance function to the hošek-wilkie skylight model. *IEEE Computer Graphics and Applications*, 33:44–52, 2013. 2

[10] W. Jakob. Mitsuba renderer, 2010. http://www.mitsuba-renderer.org. 5

[11] S. B. Knorr and D. Kurz. Real-time illumination estimation from faces for coherent rendering. In *Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, pages 113–122, 2014. 1, 2

[12] J. Lalonde, A. A. Efros, and S. G. Narasimhan. Estimating natural illumination from a single outdoor image. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 183–190, 2009. 2

[13] J.-F. Lalonde and I. A. Matthews. Lighting estimation in outdoor image collections. *International Conference on 3D Vision (3DV)*, 1:131–138, 2014. 2

[14] C. LeGendre, W.-C. Ma, G. Fyffe, J. Flynn, L. Charbonnel, J. Busch, and P. Debevec. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *ACM SIGGRAPH 2019 Talks*, 2019. 2

[15] C. Li, K. Zhou, and S. Lin. Intrinsic face image decomposition with human face priors. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 218–233, Cham, 2014. 2

[16] G. Liu, D. Ceylan, E. Yumer, J. Yang, and J. Lien. Material editing using a physically based rendering network. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 2280–2288, 2017. 1

[17] S. Lombardi and K. Nishino. Radiometric scene decomposition: Scene reflectance, illumination, and geometry from rgb-d images. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 305–313, 2016. 1

[18] R. Ramamoorthi and P. Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of SIGGRAPH 2001*, page 497–500, 2001. 2

[19] R. Ramamoorthi and P. Hanrahan. A signal-processing framework for inverse rendering. In *Proceedings of SIGGRAPH 2001*, page 117–128, 2001. 2

[20] H. Shim. Faces as light probes for relighting. *Optical Engineering*, 51(7):1 – 8, 2012. 1

[21] Z. Shu, E. Yumer, S. Hadap, K. Sunkavalli, E. Shechtman, and D. Samaras. Neural face editing with intrinsic image disentangling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5444–5453, 2017. 2

[22] G. Stratou, A. Ghosh, P. Debevec, and L. Morency. Effect of illumination on automatic expression recognition: A novel 3d relightable facial database. In *Face and Gesture 2011*, pages 611–618, 2011. 3, 4, 5, 7

[23] T. Sun, J. T. Barron, Y.-T. Tsai, Z. Xu, X. Yu, G. Fyffe, C. Rhemann, J. Busch, P. Debevec, and R. Ramamoorthi. Single image portrait relighting. *ACM Trans. Graph.*, 38(4), July 2019. 2

[24] H. Weber, D. Prévost, and J. Lalonde. Learning to estimate indoor lighting from 3d objects. In *Proceedings of the International Conference on 3D Vision (3DV)*, pages 199–207, 2018. 2

[25] R. Yi, C. Zhu, P. Tan, and S. Lin. Faces as lighting probes via unsupervised deep highlight extraction. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 317–333, September 2018. 2

[26] J. Zhang and J. Lalonde. Learning high dynamic range from outdoor panoramas. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 4529–4538, 2017. 2

[27] J. Zhang, K. Sunkavalli, Y. Hold-Geoffroy, S. Hadap, J. Eisenmann, and J.-F. Lalonde. All-weather deep outdoor lighting estimation. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 2

[28] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 7

[29] H. Zhou, S. Hadap, K. Sunkavalli, and D. W. Jacobs. Deep single-image portrait relighting. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct. 2019. 2