This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

VENI: Variational Encoder for Natural Illumination

Paul Walker[1] James A. D. Gardner[2,3] Andreea Ardelean[1] William A. P. Smith[2,3] Bernhard Egger[1]

[1]Friedrich-Alexander-Universität Erlangen-Nürnberg
[2]University of York [3]pxld.ai

https://paul-pw.github.io/veni

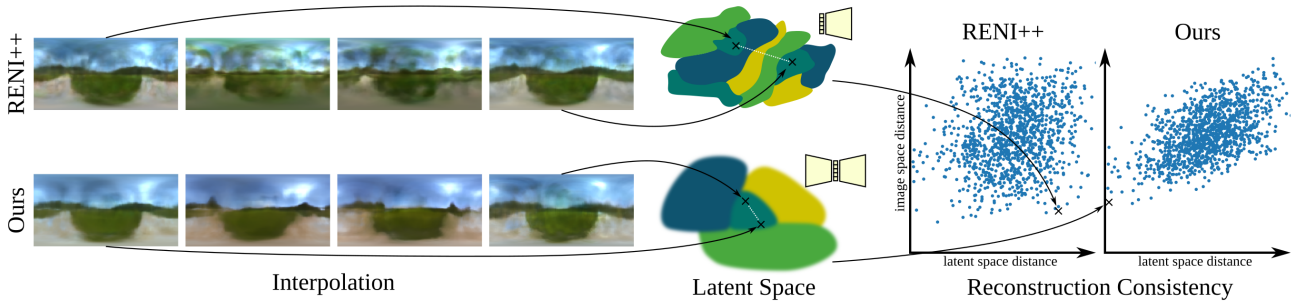

Figure 1. We build a rotation-equivariant variational autoencoder model to address the limitations of the current state-of-the-art illumination prior (RENI++). Our model produces a better-behaved latent space, which we demonstrate by evaluating its uniqueness through optimization of random latent codes and its reconstruction consistency via the correlation between latent space and image space distances.

Abstract

Inverse rendering is an ill-posed problem, but priors such as illumination priors can help simplify it. Existing work either disregards the spherical and rotation-equivariant nature of illumination environments or does not provide a well-behaved latent space. We propose a rotation-equivariant variational autoencoder that models natural illumination on the sphere without relying on 2D projections. To preserve the SO(2)-equivariance of environment maps, we use a novel Vector Neuron Vision Transformer (VN-ViT) as encoder and a rotation-equivariant conditional neural field as decoder. In the encoder, we reduce the equivariance from SO(3) to SO(2) using a novel SO(2)-equivariant fully connected layer, an extension of Vector Neurons. We show that our SO(2)-equivariant fully connected layer outperforms standard Vector Neurons when used in our SO(2)-equivariant model. Compared to previous methods, our variational autoencoder enables smoother interpolation in latent space and offers a more well-behaved latent space.

1. Introduction

Inverse rendering is ill-posed, an image can be created by multiple combinations of shape, material and lighting [5, 27]. To solve this, humans draw on strong, learned priors over the space of possible illuminations [30], especially a lighting-from-above prior [30, 39, 42], and it has been shown that humans perform better on inverse rendering tasks under natural illumination conditions, with performance degrading under artificial illumination [17]. Whilst natural illumination is complex and challenging to model, it displays statistical regularities [13], particularly in outdoor scenes. For example, the strongest sources of illumination are the sun and the skylight, which produce only a limited range of colors. Additionally, illumination environments have a canonical up direction, with any rotation about the vertical direction being equally likely.

However, instead of relying on learned priors for illumination, it is still common in inverse rendering to reconstruct the lighting directly [4, 48, 49, 57] or to rely on simple statistical models of spherical harmonics parameters of custom datasets [2, 3, 14, 52].

Recently, there has been some work to build natural illumination priors using spherical neural fields [7, 19, 20]. Such continuous representations are useful in an inverse rendering setting [21, 24] since the incident illumination can be queried from any direction. Some of these models [19, 20] exploit the vertical rotation-equivariance inherent to illumination environments, leading to a more efficient model and negating the need for rotation augmentation during training. However, they are constructed as decoder-only architectures due to the difficulty of building a rotation-equivariant encoder. For example, RENI++ [20] employs an autodecoder

architecture [31], in which latent codes are randomly initialized for each training and testing image and jointly optimized with the model. This leads to two key limitations. First, similar images are often initialized with completely different latent codes, which can cause the model to represent the same image multiple times within its latent space, *i.e*., the latent space is not unique. Second, the training cannot scale to large datasets since a latent code must be optimized per-image and the latent space uniqueness further degrades as large numbers of similar images are initialized with different latent codes.

To overcome these two limitations while retaining rotation-equivariance and the benefits of a neural field representation, we propose a novel rotation-equivariant variational autoencoder architecture for spherical signals. Following the principles of the Vector Neuron [11] framework, we design a novel vision transformer *encoder* that computes the latent codes, which are then mapped back to the image space via a neural field decoder. This enables us to use much larger datasets, as we no longer need to explicitly optimize a latent code for each image, but instead encode the image to a latent code via a forward pass through our rotation-equivariant encoder. By construction, encoding two similar images now leads to two similar latent codes, thereby addressing the weakness of RENI++ and improving the uniqueness of the latent space. This enables smooth illumination transitions through latent space interpolation.

To summarize, our contributions are:

- A rotation-equivariant panoramic vision transformer encoder enabled by a novel SO(2)-equivariant extension to Vector Neurons
- A rotation-equivariant natural illumination prior with a well-behaved, unique latent space
- Extensive evaluation of latent space uniqueness and reconstruction consistency using intuitive metrics

2. Related Work

Lighting Representations. The lighting at any point in 3D space can be represented as a spherical signal that defines the incident light intensity and color from every direction. In computer graphics, it is often assumed that the illumination environment is the same across all points in the scene, *i.e*., it models distant illumination. The illumination environment is often modeled by an environment map [34, 44], which is a projection of the spherical signal onto a 2D image via equirectangular projection or cube mapping [23]. However, both projections introduce some level of distortion, resulting in irregular sampling with respect to the original spherical image. Since available datasets are typically provided as equirectangular projected images [19, 22, 29], the associated sampling issues need to be addressed during training.

Alternative representations for illumination environments used in inverse rendering problems are Spherical Harmonics (SH) [2, 4, 14, 34, 52] and Spherical Gaussians (SG) [44, 49, 56, 57]. Since both express illumination environments as spherical signals, they neither introduce distortion nor suffer from irregular sampling. Still, Boss *et al*. [7] and Gardner *et al*. [20] have shown that neural field-based light modeling can capture more detailed environments using fewer parameters than SG or SH, so we follow their approach for our decoder.

Illumination Priors. Many inverse rendering works do not rely on illumination priors and instead reconstruct lighting directly [4, 48, 49, 57]. This often leads to unrealistic lighting estimations and hinders the estimation of other parameters, such as albedo. Since real world illumination exhibits some regularities [13], illumination priors have been shown to help constrain the inverse rendering task [21].

Recent works in light estimation focus on outpainting a 360° HDR environment map from an LDR crop to estimate the lighting in a scene [10, 47, 55]. This introduces an implicit prior on the illumination. Zhan *et al*. [55] and Dastjerdi *et al*. [10] first predict the location of the light sources from the LDR crop using SG and then use a generative adversarial network (GAN) to reconstruct the full HDR image based on the LDR crop and the light source prediction. Wang *et al*. [47] train a two branch StyleGAN [25] to generate LDR and HDR panoramas. Using this, they propose a GAN inversion method to find the latent code of an LDR crop to predict the HDR panorama. Phongthawee *et al*. [32] propose to inpaint a chrome ball into a scene to estimate the lighting instead of directly outpainting a 360° panorama. This can also be seen as an outpainting method, since the inpainted chrome ball shows what is behind the camera. While these models can reconstruct high quality illumination environments, they require that a part of the illumination environment is visible in the scene. Without which, these models cannot be used as an illumination prior.

Earlier works that do not have this limitation proposed statistical models over SH parameters of custom datasets [2, 3, 14, 52], while more recent works rely on neural approaches [7, 20, 40]. Sztrajman *et al*. [40] build an environment map convolutional autoencoder and Boss *et al*. [7] use a neural field to model pre-integrated lighting. Closest to our approach is RENI++ [20], which proposes a natural illumination prior that is rotation-equivariant by design. RENI++ uses an autodecoder architecture [31], meaning that latent codes are initialized randomly per training/testing image and then jointly optimized with the model [20]. This approach results in a latent space capable of representing an image by multiple, different latent codes, *i.e*., the latent space is not unique. This adversely affects downstream performance. Our autoencoder approach overcomes this limitation, leading to a more well-behaved latent space.

Rotation-equivariance. Lighting environments are inherently rotation-equivariant. Any rotation around the up-axis

results in another valid illumination environment. This property has been largely overlooked in priors for natural illumination [3, 10, 14, 32, 47, 55], while other works try to achieve it via data augmentation [40, 52]. rotation-equivariance in general is an important property in computer vision [8]. Some methods design rotation-equivariant models using convolutions with steerable kernels [9, 33, 41, 50]. Another approach is proposed by SE(3) Transformer [18], which introduces a rotation-equivariant attention mechanism by combining attention [45] with tensor field convolution [41]. Steerable kernels and SE(3) transformers are restricted to convolutions, which limits their applicability to non-convolutional models.

A more general framework for building SO(3)-equivariant neural networks is Vector Neurons [11]. By providing basic SO(3)-equivariant building blocks, such as linear and ReLU layers, it enables the design of a wide range of equivariant model architectures. For example, VN-Transformer [1] applies the Vector Neuron framework to transformer models. We rely on VN-Transformer to achieve rotation-equivariance in our encoder. The Vector Neurons approach was also used by RENI and RENI++ to build a rotation-equivariant prior for natural illumination [19, 20]. While RENI uses the full Gram matrix, to achieve rotation-equivariance at the cost of $O(n^2)$ complexity relative to the size of the latent space, RENI++ uses the VN-Invariant layer [11], reducing complexity to $O(n)$ [20]. We follow the approach of RENI++ to obtain a rotation-equivariant decoder.

Panoramic Vision Transformers. 360° panoramic images present inherent challenges, as the 2D projection of such data (*e.g.*, equirectangular projection, cube mapping) inevitably introduces distortion. Based on the transformer architecture [45], Vision Transformers use global self-attention to solve many computer vision tasks [12], but standard Vision Transformers are not designed to handle distortion. Shen *et al*. [37] propose the use of sphere tangent-patches to remove the negative effects of distortion when using vision transformers with 360° equirectangular projected images, while other methods address it via deformable convolutions [53, 54, 58]. Yun *et al*. [54] use deformable convolutions with fixed offsets as linear projection into the transformer encoder. Zhao *et al*. [58] propose a distortion-aware transformer block that uses deformable convolution with learnable offsets. Yuan *et al*. [53] use deformable convolutions with learnable offsets in the output projection of a 360° image segmentation model.

All of these methods address the distortion introduced when mapping the spherical 360° image to 2D, but they disregard the true spherical 3D nature of panoramic images. Consequently, all of them must explicitly account for such distortions. In contrast, we operate directly in the 3D domain and build a vision transformer upon the VN-Transformer [1]. This approach eliminates the need for distortion handling, as no projection from 3D to 2D is involved. The technique we use is closest to early fusion, as proposed in VN-Transformer [1].

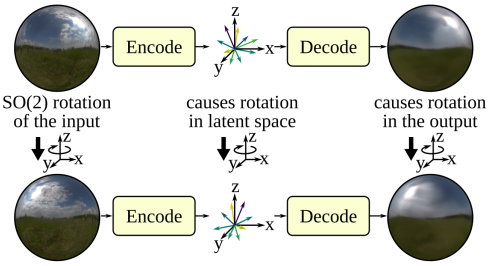

Figure 2. Our model is rotation-equivariant, meaning that a rotation of the input environment map leads to a corresponding rotation in latent space and in the reconstructed environment map.

3. Methods

Our goal is to learn a prior for natural illumination with a structured latent space, which will make operations in the latent space (*e.g.* interpolations) more semantically meaningful. Building on RENI++ [20], we address their architectural issues by implementing a rotation-equivariant variational autoencoder that lets the model implicitly learn a structured latent space. We adopt the RENI++ decoder and introduce a novel Vector Neuron Vision Transformer encoder.

SO(2)-equivariance. A model is rotation-equivariant if a rotation of the input induces the same rotation of the output, as visualized in Figure 2. For natural, outdoor illumination, where there is always a well-defined horizon line and up-axis, only rotations around the up-axis lead to other realistic illumination environments. Rotations around other axes will result in unrealistic illumination environments. Therefore, we design an SO(2)-equivariant model.

SO(2)-equivariant fully connected layer. Vector Neurons, as proposed by Deng *et al*. in [11], provide a framework for SO(3)-equivariant networks. We deal with multidimensional data (*i.e.*, direction and color), but only desire equivariance in the x and y dimensions (*i.e.*, SO(2)-equivariance). Rotations around any other axis than the up-axis, particularly rotations of the color vectors, produce unrealistic illumination environments. Therefore, these dimensions should remain invariant to rotations in the xy-plane. Since both the invariant dimensions and the equivariant dimensions contain important information, we aim for them to influence each other while preserving rotation-equivariance in the x and y dimensions and invariance in the other dimensions.

We extend Vector Neurons by combining invariant operations and equivariant operations in one neuron, resulting in SO(2)-equivariance. Given the input, $\mathbf{X} = [\mathbf{X}_{eq}, \mathbf{X}_{inv}] \in \mathbb{R}^{d_{in} \times (2+c_{inv})}$, consisting of the equivariant component, $\mathbf{X}_{eq} \in \mathbb{R}^{d_{in} \times 2}$, in our case, the x and y dimensions and the invariant component, $\mathbf{X}_{inv} \in \mathbb{R}^{d_{in} \times c_{inv}}$, where

$c_{in} = 4$ is the number of invariant input dimensions, here, the z and color dimensions, and d_{in} is the number of inputs to the fully connected layer, we arrive at the intermediate values:

$$\mathbf{T}_{inv} = [\mathbf{X}_{inv}, \mathbf{1}_{d_{in} \times 1}] \in \mathbb{R}^{d_{in} \times (c_{inv}+1)} \quad (1)$$

$$\mathbf{T}'_{inv} = [\mathbf{X}_{inv}, \|\mathbf{X}_{eq}\|] \in \mathbb{R}^{d_{in} \times (c_{inv}+1)} \quad (2)$$

where the concatenation with the column of ones in $\mathbf{T}_{inv}$ allows us to directly incorporate a bias in the bilinear combination with the equivariant inputs $\mathbf{X}_{eq}$. In $\mathbf{T}'_{inv}$, the concatenation with $\|\mathbf{X}_{eq}\|$, which is the L2 norm in the last (equivariant) dimensions, allows the invariant outputs to depend on the equivariant inputs without losing invariance. Given the weight and bias matrices,

$$\mathbf{W}_{eq} \in \mathbb{R}^{d_{out} \times (c_{inv}+1) \times d_{in}}, \mathbf{W}_{inv} \in \mathbb{R}^{d_{out} \times c_{inv} \times d_{in} \times (c_{inv}+1)}$$

$$\mathbf{B}_{inv} \in \mathbb{R}^{d_{out} \times c_{inv}},$$

we define an SO(2)-equivariant fully connected layer, $f(\mathbf{X}, \mathbf{W}_{eq}, \mathbf{W}_{inv}, \mathbf{B}_{inv})$, as follows:

$$\mathbf{Y}_{eq,o,v} = \sum_{i=1}^{d_{in}} \sum_{k=1}^{c_{inv}+1} \mathbf{W}_{eq,o,k,i} \cdot \mathbf{T}_{inv,i,k} \cdot \mathbf{X}_{eq,i,v} \quad (3)$$

$$\mathbf{Y}_{inv,o,v} = \sum_{i=1}^{d_{in}} \sum_{k=1}^{c_{inv}+1} \mathbf{W}_{inv,o,v,i,k} \cdot \mathbf{T}'_{inv,i,k} + \mathbf{B}_{inv,o,v} \quad (4)$$

$$\mathbf{Y} = f(\mathbf{X}, \mathbf{W}_{eq}, \mathbf{W}_{inv}, \mathbf{B}_{inv}) = [\mathbf{Y}_{eq}, \mathbf{Y}_{inv}] \quad (5)$$

Where $\mathbf{Y}_{eq}$ is a bilinear combination of the equivariant input $\mathbf{X}_{eq}$ and the invariant input $\mathbf{X}_{inv}$ and $\mathbf{Y}_{inv}$ is a linear combination of the invariant input $\mathbf{X}_{inv}$ and the length of the equivariant input vectors $\|\mathbf{X}_{eq}\|$. Proof of SO(2)-rotation-equivariance can be found in the supplementary material.

Vector Neuron Vision Transformer (VN-ViT). Our key contribution is a Vision Transformer and Vector Neuron-based encoder (VN-ViT), that can encode spherical images to SO(2) rotation-equivariant latent codes. Figure 3 shows an overview of the full method. The model architecture is based on vision transformers [12], but since we work with spherical images and require rotation-equivariance, we replace all ViT components with their Vector Neuron counterparts [11]. We cannot use positional encoding, since that would break equivariance. Instead, we concatenate the color values of each sample with the direction vector of where it was sampled from. Patches are then projected to the dimensions of the transformer using an SO(2)-equivariant fully connected layer. We refer to the output of the projection as patch embeddings. Following the Vision Transformer paper, we prepend an output token to the patch embeddings [12]. This token is only learnable in the invariant dimensions of our input, in the equivariant dimensions, it is set to zero. Having the token be learnable in any of the equivariant dimensions would break equivariance. The patch embeddings are then fed into a standard SO(3)-equivariant VN-Transformer [1]. The output of the transformer at the output token serves as a representation for the full image. This output is then projected from the transformer dimension to the latent dimension using two SO(2)-equivariant fully connected layers, which compute μ and σ that, following the reparameterization trick [26], yield the latent code. We found that using an SO(2)-equivariant input and output projection with a VN-Transformer encoder produces the best results.

Rotation-Equivariant Patching. We sample patches from the spherical image using our SO(2)-equivariant patching strategy, as shown in Figure 3. Patches are sampled as vertical stripes on the sphere. A rotation of the environment map around the up axis by a multiple of the stripe width results in a permutation of the color values in the patches. Since the transformer is permutation invariant, having only this without any positional encoding would make the model rotation invariant. The patches become equivariant by adding a directional component in the form of direction vectors to each pixel by concatenating the color values with the direction vectors. A rotation of the environment map around the up-axis by a multiple of the stripe width, while keeping the patches in place, is equivalent to a joint rotation of both the patches and the environment map around the up-axis. Both result in the same rotated patch embeddings. Thus, our patching strategy is SO(2)-equivariant. In practice, we use 64 patches as a trade-off between memory usage and rotation-equivariance. Combining the directional vectors with the color values, is proposed in a similar way as early fusion of non-spatial attributes in [1]. We use early fusion rather than late fusion since the direction vectors only provide directional information and do not hold any additional information.

Since we work with spherical data but sample pixels from an equirectangular projection of spherical images, we have to ensure that the sampling is regularly distributed on the sphere. We evenly sample the azimuth φ and distribute the polar angle θ as:

$$\theta = \arccos(x),\ x \in [-1, 1] \quad (6)$$

By handling spherical data directly instead of working with projected images, we do not have to account for the distortion that would otherwise be introduced by the 3D to 2D projection. Our approach of combining direction vectors and color values enables the model to know where each pixel was sampled from, eliminating the need for distortion correction.

Rotation-Equivariant Variational Sampling. Since our latent space is in 3D, variational sampling has to also take place in 3D space. The naive approach of sampling three 1D distributions results in an axis-aligned distribution, which is not rotation-equivariant. To solve this issue, we sample from a spherical normal distribution that is defined by a 3D mean

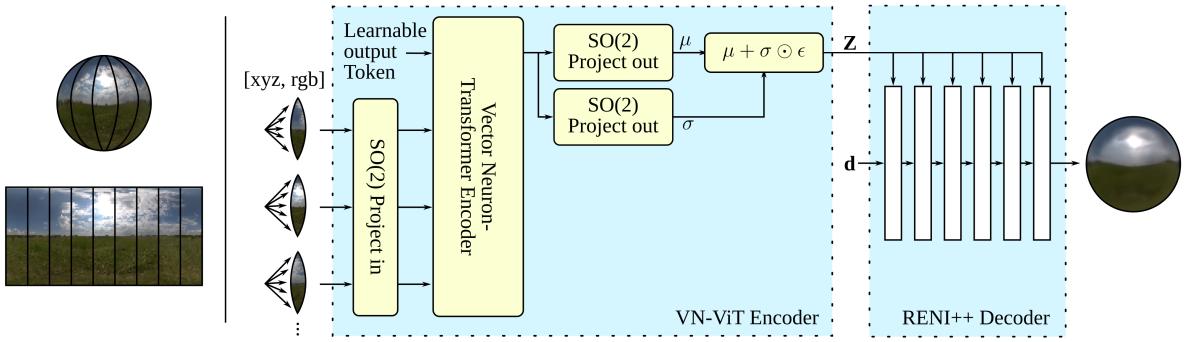

Figure 3. Model Overview: We adapt the Vision Transformer architecture to SO(2) equivariance on spherical images. Splitting the 360° spherical image into vertical stripes and embedding the direction vectors and color values of the pixels in each patch using an SO(2)-equivariant projection. Then we feed the sequence into a Vector Neuron-Transformer. We use a class token that is learnable in non-equivariant dimensions and zero in equivariant dimensions. The output is projected to μ and $\log(\sigma^2)$ and reparameterized. A rotation of the input by a multiple of the patch width results in the output of the encoder being rotated in the same way. The output latent code of the encoder is fed into the RENI++ decoder.

vector and a 1D variance. This distribution is isotropic and thus rotation-equivariant.

RENI++ decoder. We adopt the same decoder as proposed in RENI++ [20], which introduces a rotation-equivariant conditional neural field autodecoder architecture that leverages attention for conditioning. This design is motivated by findings that attention-based conditioning of neural fields outperforms alternative methods of conditioning [35]. The input to the RENI++ neural field $\mathcal{D}$ is a direction vector $\mathbf{d}$ and the 3D latent vector $\mathbf{Z}$. The output is the color $\mathbf{c}$ at direction $\mathbf{d}$. To achieve rotation-equivariance, $\mathbf{d}$ and $\mathbf{Z}$ are encoded so they are invariant to simultaneous rotations, so $\mathcal{D}(\mathbf{Rd}, \mathbf{RZ}) = \mathcal{D}(\mathbf{d}, \mathbf{Z})$ with $\mathbf{R} \in SO(3)$. This makes the neural field equivariant to rotations of $\mathbf{Z}$ only. Rotating $\mathbf{Z}$ is equivalent to rotating the spherical signals such that $\mathcal{D}(\mathbf{d}, \mathbf{RZ}) = \mathcal{D}(\mathbf{R}^\top \mathbf{d}, \mathbf{Z})$. The invariant encoding of $\mathbf{d}$ and $\mathbf{Z}$ to simultaneous rotations is achieved by encoding the direction $\mathbf{d}$ relative to the latent code: $\mathbf{d}' = \mathbf{Z}^\top \mathbf{d}$ and applying the Vector Neuron invariant layer [11] to the latent code: $\mathbf{Z}' = \text{VN-Inv}(\mathbf{Z})$.

Pretraining. For rendering, the full dynamic range of natural light is essential. To let our model learn this full range, our dataset must also consist of images that cover the entire dynamic range of natural light and not clipping the brightest parts of the image. We use two datasets to train our model. One is the RENI++ dataset, consisting of 1,694 HDR equirectangular images obtained under a CC0 1.0 Universal Public Domain Dedication license [20]. It features a wide variety of outdoor scenes and illumination conditions. These images have a resolution of 128×64. This low resolution is sufficient for our purposes. The RENI++ dataset is limited in size and further availability of 360° HDR outdoor illumination environments is limited. However, 360° LDR data is readily available, for example in the form of Google street view data. Although we cannot use LDR data directly, there exists a large body of work that focuses on reconstructing HDR data from LDR images [16, 38, 46, 51]. Some of which even take 360° environment maps into account [38, 51]. In addition to the RENI++ dataset, we also use the streetlearn dataset [29], consisting of a large number of LDR 360° images captured by Google street view in Manhattan. We convert 43,310 equirectangular 360° images from the Manhattan part of the streetlearn dataset to HDR using [51]. The converted streetlearn dataset is of lower quality than the RENI++ dataset since the HDR reconstruction process is imperfect and the dataset predominantly contains cityscapes, whereas we want good HDR reconstruction of varied outdoor scenes. We address this in our training curriculum by first pretraining our model on the much larger streetlearn dataset and then finetuning it on the RENI++ dataset.

4. Experiments

Losses. To better match how the human visual system reacts to luminance, we train our model in log space. This is common practice for models that deal with HDR data [16, 38, 46, 51]. With a loss in linear space, areas with high luminance would overpower important features in areas with low or medium luminance. A loss in log space alleviates this issue by spreading the loss approximately linearly across the perceived luminance range [16].

Additionally, there is scale ambiguity within our dataset, because the HDR images are captured at unknown exposure values (EV). We only know the relative luminance between pixels, not the absolute luminance of each pixel. Any image multiplied by a positive scale factor k, would still be valid.

Our dataset contains high frequency details, which are especially prevalent at the boundary between sun and sky,

where there is a large change in intensity in a small area. To enable our model to capture more high frequency detail, we use a loss from depth prediction, the Mean Absolute Gradient Error (MAGE) [6]. This loss improves high frequency detail by operating on the gradients of the image. High frequency details will lead to high gradients and thus the error in high frequency areas is more strongly penalized. During training, we decode the full equirectangular images. Thus, we have to account for the irregular sampling by weighting with the sin of the polar angle of each pixel, $\sin(\theta_i)$:

$$\mathcal{L}_{MAGE} = \frac{1}{M}\sum_{j=1}^{M}\frac{1}{N_j}\sum_{i=1}^{N_j}\sin(\theta_i)|\nabla_S f(\mathbf{I})_i^j - \nabla_S \mathbf{I}_i^j|, \quad (7)$$

with ∇_S denoting the spatial derivative operator Scharr (S) [36] and $M = 2$ denoting the number of scales, used for the multi-scale derivative loss. The inputs to the loss are $\mathbf{I}$, the ground truth image in log space, and $f(\mathbf{I})$, the model output in log space. Since this loss is applied in log space, it is scale invariant. In log space, Scharr removes the influence of any arbitrary scale factor k.

We also follow the approach of RENI++ and LANet [20, 51] and use another technique from depth prediction [15, 28], the scale invariant loss. This loss calculates the relative error over the entire image, rather than on the gradients of the image:

$$\mathcal{L}_{scale\text{-}inv} = \frac{1}{N}\sum_{i=1}^{N}(R_i)^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N}R_i\right)^2. \quad (8)$$

where $R_i = \sin(\theta_i)(f(\mathbf{I})_i - \mathbf{I}_i)$ at pixel i. This helps us learn a scale invariant representation of the images. To encourage our model to produce accurate color representations, we again follow the approach of RENI++ [20] and use a cosine similarity loss. Since this loss operates on the direction of RGB vectors, it is also scale invariant:

$$\mathcal{L}_{cosine} = 1 - \frac{1}{N}\sum_{i=1}^{N}\sin(\theta_i)\frac{f(\mathbf{I})_i \cdot \mathbf{I}_i}{\|f(\mathbf{I})_i\|\|\mathbf{I}_i\|}. \quad (9)$$

We train our model using variational sampling and regularize the latent space to follow a standard normal distribution using Kullback Leibler divergence (KLD). Since we use a 3D isotropic distribution for our variational sampling, we must also adjust the KLD loss term accordingly:

$$\mathcal{L}_{KLD} = \frac{1}{D}\sum_{i}^{D} -0.5\cdot(3+3\cdot\log(\sigma_i) - \|\mu_i\|^2 - 3\cdot\sigma_i), \quad (10)$$

where D is the latent dimension. Our full training loss is:

$$\mathcal{L} = 0.5\cdot\mathcal{L}_{MAGE} + \mathcal{L}_{scale\text{-}inv} + \mathcal{L}_{cosine} + 0.01\cdot\mathcal{L}_{KLD}. \quad (11)$$

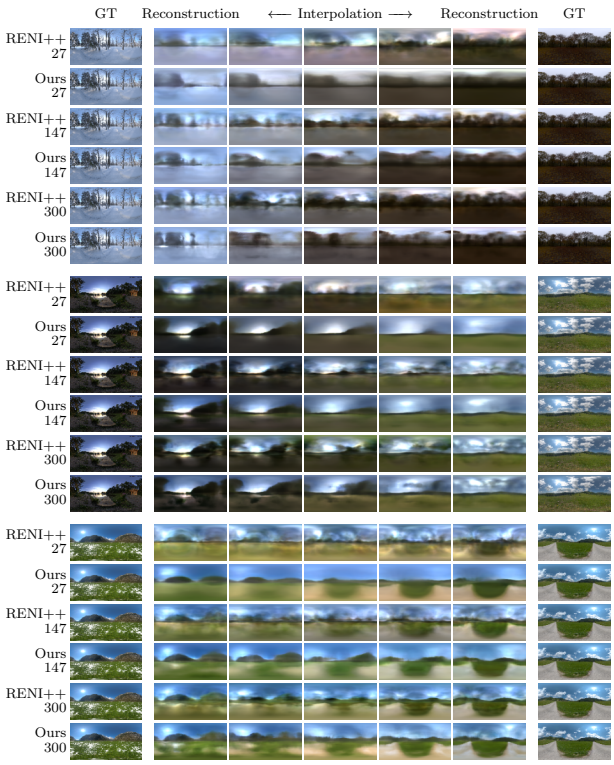

Figure 4. Interpolations using our model (with direct latent optimization) and RENI++ with different latent sizes. Three image pairs are interpolated: snow-forest, lake-field, mountain-road.

Quality Metrics. We report in Table 1, the PSNR, SSIM and LPIPS scores in LDR tone-mapped space, as well as PSNR in linear HDR space, all averaged across the test set. RENI++ results are taken from their paper, as we found them reproducible. For our model, we report metrics for both the full autoencoder pass and decoder-only latent optimization, the latter being more comparable to RENI++ which supports only latent optimization. Following RENI++, we also optimize a per-image scale factor k during latent optimization. This factor scales the decoder output and is not considered in the full autoencoder pass. Especially in the low-dimensional case of $D = 27$, our model strongly outperforms RENI++. We include a qualitative comparison of the three approaches in the supplementary material.

Interpolation. A well-behaved latent space enables smooth interpolations between latent codes. Each step of the interpolation should correspond to a realistic and semantically meaningful illumination environment. For example, interpolating between a sunrise and a midday illumination environment should result in the sun moving smoothly across the sky and the color temperature changing gradually from daylight to sunset without introducing artifacts.

Figure 4 shows interpolations of RENI++ and our model for the same image pairs. Our model meaningfully interpo-

	RENI++ (optimization)				Ours (AE)				Ours (optimization)			
D	PSNR↑	SSIM↑	LPIPS↓	PSNR HDR↑	PSNR↑	SSIM↑	LPIPS↓	PSNR HDR↑	PSNR↑	SSIM↑	LPIPS↓	PSNR HDR↑
27	18.02	0.39	0.62	33.00	18.78	0.46	0.62	33.37	**20.33**	**0.51**	**0.61**	**33.92**
147	21.13	0.51	**0.55**	34.30	19.40	0.48	0.59	33.59	**21.89**	**0.54**	**0.55**	**35.07**
300	22.10	0.55	0.52	35.10	19.47	0.48	0.59	33.68	**22.68**	**0.57**	**0.48**	**35.59**

Table 1. Comparison of reconstruction quality metrics between RENI++ and our model (autoencoder pass and direct latent code optimization), across varying latent dimensions D. PSNR, SSIM and LPIPS are in LDR tone-mapped space, PSNR HDR is in linear HDR space.

	Uniqueness↓		Reconstruction Consistency↑	
D	RENI++	Ours	RENI++	Ours
27	1.46	**0.04**	0.17	**0.50**
147	1.11	**0.43**	0.12	**0.30**
300	1.03	**0.57**	0.07	**0.23**

Table 2. Uniqueness is the mean squared error (MSE) between an image produced by an optimized latent code and an image produced by the interpolation midpoint of two latent codes optimized to the same image. Reconstruction Consistency are the Spearman correlation coefficients between the MSE of random latent pairs and the MSE between their corresponding image reconstructions.

Dataset size	RENI++	Ours (AE)	Ours (optimized)
1500	20.11	16.00	**20.99**
43260	17.14	16.77	**19.90**

Table 3. Mean PSNR of our model and the RENI++ model trained on a large and a small subset of the HDR converted streetlearn dataset, evaluated on the same test set from the converted streetlearn dataset.

lates between the start and target image, whereas RENI++ introduces artifacts in the interpolation steps. In the snow-forest example, our model shows a smooth color change, while RENI++ introduces a sun in the middle of the interpolation, that is not present in either the source or the target image. Our model can also smoothly interpolate between sunrise and midday illumination environments, as seen in the lake-field example. The yellow reflection in the lake smoothly disappears and the scene transitions to a more bluish midday illumination environment. In the mountain-road interpolation, RENI++ introduces artifacts and noise in the sky during interpolation, whereas our model does not. This is especially visible with $D = 27$.

Uniqueness. One goal of our model is uniqueness of the latent space. To test this, we optimize two random latent codes sampled from a normal distribution $\mathcal{N}(0, 1)$ to fit the same image. If the latent space is unique, no other images should be represented between the two optimized latent codes. We measure model uniqueness by interpolating between two optimized latent codes, as shown in Figure 5. For unique models, the output images do not diverge during interpolation. Since the output images barely diverge for our model and diverge significantly for RENI++, we can conclude that our model is more unique than RENI++. To objectively measure uniqueness, we compute the MSE between the image produced by the first optimized latent code and the image produced by the latent interpolation midpoint. Higher error indicates that the model is less unique. Table 2 shows this metric for our model and the RENI++ model over various latent dimensions D. We include further examples of the interpolation between two optimized latent codes in the supplementary material.

The RENI++ autodecoder architecture hinders uniqueness in the model. Two similar images may be initialized with vastly different latent codes, while our model encodes two similar images to similar latent codes by using a variational autoencoder architecture. With the small dataset RENI++ uses for training, non-uniqueness is not a significant issue. However, with a larger dataset, there are more similar images that are then initialized with different latent codes, decreasing the uniqueness of the model even further. This also leads to a decrease in RENI++ performance as dataset size increases, as seen in Table 3. While RENI++ performance degrades with an increased dataset size, our model's performance improves with an autoencoder pass and does not degrade as much with optimization.

We also measure the reconstruction consistency of our model (see Table 2). This metric calculates the correlation coefficients between the error of random latent pairs and the error between their corresponding image reconstructions. Following the idea of the latent-data distance constraint by Tran *et al.* [43], it tells us how well latent distance matches perceived distance and is an indicator for latent space uniqueness. A higher correlation indicates more uniqueness because a low latent error combined with a low reconstruction error (*i.e.*, a more unique model) results in a higher correlation. In contrast, a high latent error combined with a low reconstruction error (*i.e.*, a non-unique model) results in a lower correlation. The improved reconstruction quality and more unique latent space of our model enable improve-

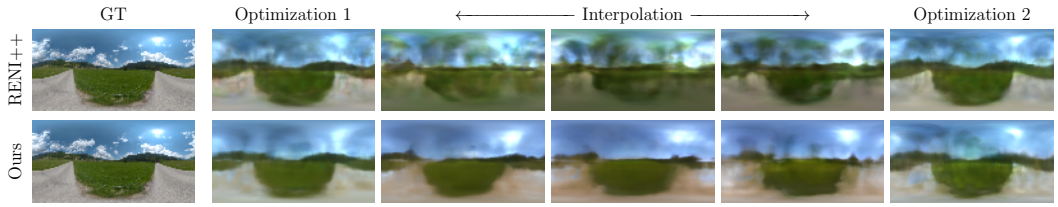

Figure 5. Interpolating between two optimized latent codes. With our model, the output images remain consistent during interpolation, suggesting that the latent space is unique. With RENI++ the output images diverge from the optimized images during interpolation, suggesting that the latent space is not unique.

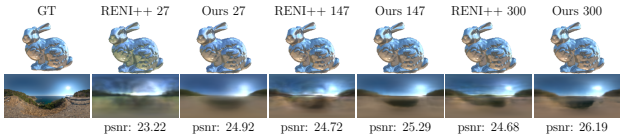

Figure 6. Application: the latent space of our model is more amenable to inverse rendering optimization, outperforming the RENI++ baseline.

Component	27	147	300
Ours	**18.78**	**19.40**	**19.47**
- scale-inv & MAGE loss	17.32	18.47	18.80
- pretraining on streetlearn	17.37	18.07	17.88
- SO(2) linear projection	16.89	17.49	17.30

Table 4. Mean PSNR on various ablations of our model

D	SO(2) projections	Full SO(2)	Full VN
27	**18.78**	17.89	17.74
147	**19.40**	18.18	18.80
300	**19.47**	17.93	18.60

Table 5. Mean PSNR comparing models only using SO(2) linear layers in the projection layers, using SO(2) layers in the entire encoder including the transformer and using VN-layers in the entire encoder.

ments in downstream applications such as inverse rendering optimization, as shown in Figure 6.

Ablations. To show that our SO(2)-equivariant extension to Vector Neurons improves upon the default SO(3)-equivariant Vector Neurons, we do an ablation study. We compare the performance of Vector Neurons with that of our SO(2)-equivariant extension in the projection layers and the full encoder across a range of latent dimensions D using the mean PSNR. For the SO(2)-equivariant transformer, we have to account for the SO(2)-equivariant layers using the invariant axes as additional learnable parameters, by decreasing the inner dimension of the SO(2)-equivariant transformer accordingly. Table 5 shows the results of this evaluation. The best results are achieved when the SO(2)-equivariant fully connected layers are used only as projection layers. Thus, to benefit from reducing the equivariance guarantee from SO(3) to SO(2), it is not necessary to reduce equivariance for the entire model. Reducing the equivariance by using SO(2)-equivariant fully connected layers as projection layers is sufficient.

We also test the losses used over MSE, as well as our pretraining method that uses the converted streetlearn dataset. Table 4 shows the results for the full autoencoder pass, comparing mean LDR PSNR over various ablations and latent dimensions. We include a qualitative comparison of the ablations in the supplementary material.

5. Conclusion

Our model produces realistic HDR environment maps with a well-behaved latent space that enables smooth interpolation. We design a rotation-equivariant vision transformer (VN-ViT) that operates on 360° panoramic images and follows the Vector Neuron principles. We introduce an SO(2)-equivariant extension to Vector Neurons and use it to reduce the equivariance of the VN-ViT from SO(3) to SO(2), resulting in an increased performance. To further enhance our model's performance, we adapt losses from depth prediction. Our model outperforms RENI++ in terms of reconstruction quality and latent-space structure. It provides a more unique latent representation and scales effectively to larger datasets, whereas RENI++ performance degrades, our model continues to improve. In conclusion, our work advances the modeling of natural illumination, and we believe this contribution will benefit downstream tasks such as inverse rendering, relighting and other vision problems that rely on accurate illumination priors.

Acknowledgments. The authors gratefully acknowledge the scientific support and HPC resources provided by the Erlangen National High Performance Computing Center (NHR@FAU) of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU). The hardware is funded by the German Research Foundation (DFG). James Gardner was supported by the EPSRC Centre for Doctoral Training in Intelligent Games & Games Intelligence (IGGI) (EP/S022325/1).

References

[1] Serge Assaad, Carlton Downey, Rami Al-Rfou', Nigamaa Nayakanti, and Benjamin Sapp. Vn-transformer: Rotation-equivariant attention for vector neurons. *Trans. Mach. Learn. Res.*, 2023, 2023. 3, 4

[2] Jonathan T. Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *CVPR*, 2013. 1, 2

[3] Jonathan T. Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE TPAMI*, 37(8):1670–1687, 2015. 1, 2, 3

[4] R. Basri and D.W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE TPAMI*, 25(2):218–233, 2003. 1, 2

[5] Peter N Belhumeur, David J Kriegman, and Alan L Yuille. The bas-relief ambiguity. *IJCV*, 35(1):33–44, 1999. 1

[6] Alexey Bochkovskiy, Amaël Delaunoy, Hugo Germain, Marcel Santos, Yichao Zhou, Stephan Richter, and Vladlen Koltun. Depth pro: Sharp monocular metric depth in less than a second, 2025. 6

[7] Mark Boss, Varun Jampani, Raphael Braun, Ce Liu, Jonathan Barron, and Hendrik PA Lensch. Neural-pil: Neural pre-integrated lighting for reflectance decomposition. In *NeurIPS*, pages 10691–10704. Curran Associates, Inc., 2021. 1, 2

[8] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478, 2021. 3

[9] Taco S. Cohen and Max Welling. Steerable CNNs. In *ICLR*, 2017. 3

[10] Mohammad Reza Karimi Dastjerdi, Jonathan Eisenmann, Yannick Hold-Geoffroy, and Jean-François Lalonde. Everlight: Indoor-outdoor editable hdr lighting estimation. In *ICCV*, pages 7420–7429, 2023. 2, 3

[11] Congyue Deng, Or Litany, Yueqi Duan, Adrien Poulenard, Andrea Tagliasacchi, and Leonidas J. Guibas. Vector neurons: A general framework for so(3)-equivariant networks. In *ICCV*, pages 12200–12209, 2021. 2, 3, 4, 5

[12] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 3, 4

[13] Ron O. Dror, Alan S. Willsky, and Edward H. Adelson. Statistical characterization of real-world illumination. *J. Vis.*, 4 (9):11–11, 2004. 1, 2

[14] Bernhard Egger, Sandro Schönborn, Andreas Schneider, Adam Kortylewski, Andreas Morel-Forster, Clemens Blumer, and Thomas Vetter. Occlusion-aware 3d morphable models and an illumination prior for face image analysis. *IJCV*, 126 (12):1269–1287, 2018. 1, 2, 3

[15] David Eigen, Christian Puhrsch, and Rob Fergus. Depth map prediction from a single image using a multi-scale deep network, 2014. 6

[16] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K. Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM TOG*, 36(6), 2017. 5

[17] Roland W Fleming, Ron O Dror, and Edward H Adelson. Real-world illumination and the perception of surface reflectance properties. *J. Vis.*, 3(5):3–3, 2003. 1

[18] Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. In *NeurIPS*, pages 1970–1981. Curran Associates, Inc., 2020. 3

[19] James Gardner, Bernhard Egger, and William Smith. Rotation-equivariant conditional spherical neural fields for learning a natural illumination prior. In *NeurIPS*, pages 26309–26323. Curran Associates, Inc., 2022. 1, 2, 3

[20] James A. D. Gardner, Bernhard Egger, and William A. P. Smith. Reni++ a rotation-equivariant, scale-invariant, natural illumination prior, 2023. 1, 2, 3, 5, 6

[21] James A. D. Gardner, Evgenii Kashin, Bernhard Egger, and William A. P. Smith. The sky's the limit: Relightable outdoor scenes via a sky-pixel constrained illumination prior and outside-in visibility. In *ECCV*, pages 126–143, Cham, 2025. Springer Nature Switzerland. 1, 2

[22] Marc-André Gardner, Kalyan Sunkavalli, Ersin Yumer, Xiaohui Shen, Emiliano Gambaretto, Christian Gagné, and Jean-François Lalonde. Learning to predict indoor illumination from a single image. *ACM TOG*, 36(6), 2017. 2

[23] Ned Greene. Environment mapping and other applications of world projections. *IEEE Comput. Graph. Appl.*, 6(11):21–29, 1986. 2

[24] Zixuan Huang, Mark Boss, Aaryaman Vasishta, James M. Rehg, and Varun Jampani. Spar3d: Stable point-aware reconstruction of 3d objects from single images. In *CVPR*, pages 16860–16870, 2025. 1

[25] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, pages 4401–4410, 2019. 2

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 4

[27] Georgios Kouros, Minye Wu, Sushruth Nagesh, Xianling Zhang, and Tinne Tuytelaars. Unveiling the ambiguity in neural inverse rendering: A parameter compensation analysis. In *CVPR Workshops*, pages 2832–2841, 2024. 1

[28] Zhengqi Li and Noah Snavely. Megadepth: Learning single-view depth prediction from internet photos. In *CVPR*, 2018. 6

[29] Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, and Raia Hadsell. The streetlearn environment and dataset, 2019. 2, 5

[30] Richard F Murray and Wendy J Adams. Visual perception and natural illumination. *Curr. Opin. Behav. Sci.*, 30:48–54, 2019. Visual perception. 1

[31] Jeong Joon Park, Peter Florence, Julian Straub, Richard Newcombe, and Steven Lovegrove. Deepsdf: Learning continuous signed distance functions for shape representation. In *CVPR*, pages 165–174, 2019. 2

[32] Pakkapon Phongthawee, Worameth Chinchuthakun, Nontaphat Sinsunthithet, Varun Jampani, Amit Raj, Pramook Khungurn, and Supasorn Suwajanakorn. Diffusionlight:

Light probes for free by painting a chrome ball. In *CVPR*, pages 98–108, 2024. 2, 3

[33] Adrien Poulenard and Leonidas J Guibas. A functional approach to rotation equivariant non-linearities for tensor field networks. In *CVPR*, pages 13174–13183, 2021. 3

[34] Ravi Ramamoorthi and Pat Hanrahan. An efficient representation for irradiance environment maps. In *Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques*, page 497–500, New York, NY, USA, 2001. Association for Computing Machinery. 2

[35] Daniel Rebain, Mark J. Matthews, Kwang Moo Yi, Gopal Sharma, Dmitry Lagun, and Andrea Tagliasacchi. Attention beats concatenation for conditioning neural fields, 2023. 5

[36] Hanno Scharr, Stefan Körkel, and Bernd Jähne. Numerische isotropieoptimierung von fir-filtern mittels querglättung. In *Mustererkennung 1997*, pages 367–374, Berlin, Heidelberg, 1997. Springer Berlin Heidelberg. 6

[37] Zhijie Shen, Chunyu Lin, Kang Liao, Lang Nie, Zishuo Zheng, and Yao Zhao. Panoformer: Panorama transformer for indoor 360° depth estimation. In *ECCV*, pages 195–211, Cham, 2022. Springer Nature Switzerland. 3

[38] Gyeongik Shin, Kyeongmin Yu, Mpabulungi Mark, and Hyunki Hong. Hdr map reconstruction from a single ldr sky panoramic image for outdoor illumination estimation. *IEEE Access*, 11:17359–17374, 2023. 5

[39] J.V. Stone, I.S. Kerrigan, and J. Porrill. Where is the light? bayesian perceptual priors for lighting direction. *Proceedings of the Royal Society B: Biological Sciences*, 276(1663):1797–1804, 2009. 1

[40] Alejandro Sztrajman, Alexandros Neophytou, Tim Weyrich, and Eric Sommerlade. High-dynamic-range lighting estimation from face portraits. In *International Conference on 3D Vision*, pages 355–363, 2020. 2, 3

[41] Nathaniel Thomas, Tess Smidt, Steven Kearnes, Lusann Yang, Li Li, Kai Kohlhoff, and Patrick Riley. Tensor field networks: Rotation- and translation-equivariant neural networks for 3d point clouds, 2018. 3

[42] Rhiannon Thomas, Marko Nardini, and Denis Mareschal. Interactions between "light-from-above" and convexity priors in visual development. *J. Vis.*, 10(8):6–6, 2010. 1

[43] Ngoc-Trung Tran, Tuan-Anh Bui, and Ngai-Man Cheung. Dist-gan: An improved gan using distance constraints. In *ECCV*, 2018. 7

[44] Yu-Ting Tsai and Zen-Chung Shih. All-frequency precomputed radiance transfer using spherical radial basis functions and clustered tensor approximation. *ACM TOG*, 25 (3):967–976, 2006. 2

[45] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*. Curran Associates, Inc., 2017. 3

[46] Chao Wang, Zhihao Xia, Thomas Leimkuhler, Karol Myszkowski, and Xuaner Zhang. Lediff: Latent exposure diffusion for hdr generation. In *CVPR*, pages 453–464, 2025. 5

[47] Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. Stylelight: Hdr panorama generation for lighting estimation and editing. In *ECCV*, pages 477–492, Cham, 2022. Springer Nature Switzerland. 2, 3

[48] Lezhong Wang, Duc Minh Tran, Ruiqi Cui, Thomson TG, Anders Bjorholm Dahl, Siavash Arjomand Bigdeli, Jeppe Revall Frisvad, and Manmohan Chandraker. Materialist: Physically based editing using single-image inverse rendering. *arXiv preprint arXiv:2501.03717*, 2025. 1, 2

[49] Zian Wang, Jonah Philion, Sanja Fidler, and Jan Kautz. Learning indoor inverse rendering with 3d spatially-varying lighting. In *ICCV*, pages 12538–12547, 2021. 1, 2

[50] Maurice Weiler, Mario Geiger, Max Welling, Wouter Boomsma, and Taco S Cohen. 3d steerable cnns: Learning rotationally equivariant features in volumetric data. In *NeurIPS*. Curran Associates, Inc., 2018. 3

[51] Hanning Yu, Wentao Liu, Chengjiang Long, Bo Dong, Qin Zou, and Chunxia Xiao. Luminance attentive networks for hdr image and panorama reconstruction. *Computer Graphics Forum*, 40(7):181–192, 2021. 5, 6

[52] Ye Yu and William A. P. Smith. Outdoor inverse rendering from a single image using multiview self-supervision. *IEEE TPAMI*, 44(7):3659–3675, 2022. 1, 2, 3

[53] Zheng Yuan, Junhua Wang, Yuxin Lv, Ding Wang, and Yi Fang. Laformer: Vision transformer for panoramic image semantic segmentation. *IEEE Signal Processing Letters*, 30: 1792–1796, 2023. 3

[54] Heeseung Yun, Sehun Lee, and Gunhee Kim. Panoramic vision transformer for saliency detection in 360° videos. In *ECCV*, pages 422–439, Cham, 2022. Springer Nature Switzerland. 3

[55] Fangneng Zhan, Changgong Zhang, Yingchen Yu, Yuan Chang, Shijian Lu, Feiying Ma, and Xuansong Xie. Emlight: Lighting estimation via spherical distribution approximation. *AAAI*, 35(4):3287–3295, 2021. 2, 3

[56] Kai Zhang, Fujun Luan, Qianqian Wang, Kavita Bala, and Noah Snavely. Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *CVPR*, pages 5453–5462, 2021. 2

[57] Yuanqing Zhang, Jiaming Sun, Xingyi He, Huan Fu, Rongfei Jia, and Xiaowei Zhou. Modeling indirect illumination for inverse rendering. In *CVPR*, pages 18643–18652, 2022. 1, 2

[58] Yinjie Zhao, Lichen Zhao, Qian Yu, Lu Sheng, Jing Zhang, and Dong Xu. Distortion-aware transformer in 360° salient object detection. In *ACM MM*, page 499–508, New York, NY, USA, 2023. Association for Computing Machinery. 3