IRISformer: Dense Vision Transformers for Single-Image Inverse Rendering in Indoor Scenes

Rui Zhu[1] Zhengqin Li[1] Janarbek Matai[2] Fatih Porikli[2] Manmohan Chandraker[1]
[1]UC San Diego [2]Qualcomm AI Research
{rzhu,zhl378,mkchandraker}@eng.ucsd.edu {jmatai,fporikli}@qti.qualcomm.com

Abstract

Indoor scenes exhibit significant appearance variations due to myriad interactions between arbitrarily diverse object shapes, spatially-changing materials, and complex lighting. Shadows, highlights, and inter-reflections caused by visible and invisible light sources require reasoning about long-range interactions for inverse rendering, which seeks to recover the components of image formation, namely, shape, material, and lighting. In this work, our intuition is that the long-range attention learned by transformer architectures is ideally suited to solve longstanding challenges in single-image inverse rendering. We demonstrate with a specific instantiation of a dense vision transformer, IRISformer, that excels at both single-task and multi-task reasoning required for inverse rendering. Specifically, we propose a transformer architecture to simultaneously estimate depths, normals, spatially-varying albedo, roughness and lighting from a single image of an indoor scene. Our extensive evaluations on benchmark datasets demonstrate state-of-the-art results on each of the above tasks, enabling applications like object insertion and material editing in a single unconstrained real image, with greater photorealism than prior works. Code and data are publicly released.[1]

1. Introduction

Inverse rendering has long been of great interest to the computer vision community owing to its promise to decompose a scene into the intrinsic factors of shape, complex spatially-varying lighting, and material, thereby enabling downstream tasks of virtual object insertion, material editing, and relighting. The problem is particularly challenging for indoor scenes, where complex appearances stem from multiple interactions among the above intrinsic factors, such as shadows, specularities, and interreflections.

Recent advances in inverse rendering has led to the emergence of numerous works undertaking either some specific aspects of this challenge (geometry [8, 24], albedo [4, 20, 33], lighting [9, 11, 18, 36]), or joint estimation [2, 19, 33, 39]. However, the task of scene decomposition can be extremely ill-posed due to the inherent ambiguity between complex lighting, geometry, and material which jointly govern image formation in indoor scenes. For example, high-intensity pixel values can be explained by either specular or light-colored material, particular local geometry, bright lighting, or by a combination of all those factors. The problem is especially severe with only a single image as input, where prior knowledge is necessary to disambiguate among all possible intrinsic decompositions that explain the image. Classical methods leverage strong heuristic priors in an optimization objective [2, 3], which may not always hold for real world scenes with complex geometry or lighting conditions.

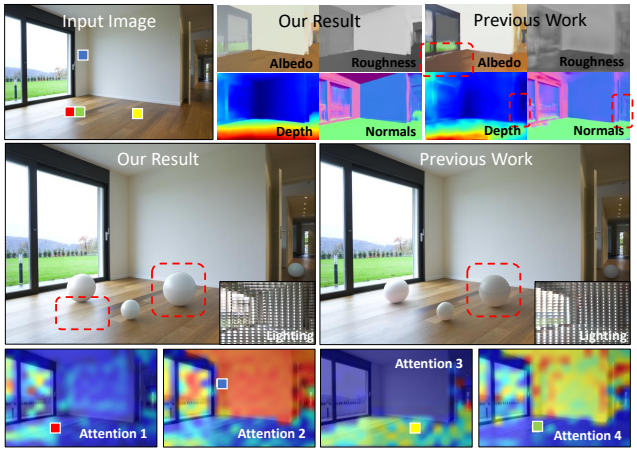

Figure 1. Given a single real world image, IRISformer simultaneously infers material (albedo and roughness), geometry (depth and normals), and spatially-varying lighting of the scene. The estimation enables virtual object insertion where we demonstrate high-quality photorealistic renderings in challenging lighting conditions compared to previous work [19]. The learned attention is also visualized for selected patches, indicating benefits of global attention to reason about distant interactions (see text for details).

[1] https://github.com/ViLab-UCSD/IRISformer

The widespread use of convolutional neural networks (CNN) and large-scale datasets for scene decomposition [4, 23, 33, 34] promotes supervised training of end-to-end multi-task models [19, 33] for joint estimation. CNN-based models have demonstrated impressive progress on inverse rendering of real world images [19, 33, 36, 39]. Nonetheless, receptive fields in CNN architectures remain largely local throughout the consecutive layers, limiting the ability to capture long-range interactions between scene elements. As shown in Fig. 1, CNN-based approaches fail to handle scenes where strong shadows or highlights abound due to complex light transport. This indicates that long-range dependencies across the image space must be exploited to provide globally coherent predictions in inverse rendering. Recently, vision transformers [7, 43] (ViT) have emerged for multiple computer vision tasks, benefiting from global reasoning via spatial attention mechanisms. In particular, dense vision transformers [25, 28, 38] are well-suited for dense prediction, which we posit can benefit inverse rendering.

In this paper, we propose to leverage vision transformers to better account for complex light transport in inverse rendering. Consider Fig. 1 as an example, where we compare our proposed transformer-based approach, *i.e*. **IRISformer** (Trans*former* for *I*nverse *R*endering in *I*ndoor *S*cenes), with a CNN-based prior state-of-the-art [19]. Note the improvement in material consistency and geometry of the floor where complex lighting governs appearance; as a result of which, the leftmost sphere is properly reflected on the floor. Additionally, IRISformer better captures global ambient lighting so that the third sphere from left is better illuminated. We also visualize the heatmaps of four patch locations shown by colored squares from selected transformer layers and heads (warmer colors indicate higher attention). By attending to large *global* regions with semantic meaning, the transformer can better disambiguate geometry material and lighting (yellow). Long-range interactions among such regions can help reason about inter-reflections (green), directional highlights (red), or shadows (blue). as well as the long-range attention to/within those homogeneous regions, the model manages to better resolve the albedo-lighting ambiguity, and making more consistent estimations.

We demonstrate that by the insightful design of single-task and multi-task models for inverse rendering with dense vision transformers, we can achieve state-of-the-art, high-quality, and globally coherent BRDF, geometry, and lighting prediction. In addition, downstream tasks like object insertion and material editing greatly benefit from our improvements, especially in scenarios of complex highlights or shadows. We achieve state-of-the-art results on all sub-tasks on real world datasets of IIW [4] and NYUv2 [34], as well as object insertion tasks compared to prior works.

Our contributions are threefold. (1) We propose the first dense vision transformer-based framework for inverse rendering in a multi-task setting. (2) We demonstrate that appropriate design choices lead to better handling of global interactions between scene components, leading to better disambiguation of shape, material and lighting. (3) We demonstrate state-of-the-art results on all tasks and enable high-quality applications in augmented reality.

2. Related Work

Inverse rendering for indoor scenes. Several prior works study the interaction of shape, material and light to estimate shape-from-shading [15, 46], intrinsic image decomposition [13, 20], material properties [1, 6, 21, 22, 32], or illumination [3, 10, 11, 44], while inverse rendering seeks to estimate all those factors simultaneously [26]. Classical methods for inverse rendering are typically posed as energy minimization with heuristic priors, for example, SIRFS [2,3] where intrinsic properties are jointly optimized with a statistical cost function. Despite early success, such models usually do not generalize well to real images with diverse appearances. With advances in deep learning, CNN-based methods have been developed to learn a generalizable model in a data-driven fashion. The recent method of NIR [33] is pre-trained with weak labels and finetuned on real images with re-rendering loss. Methods like Lighthouse [36] learn a volumetric lighting representation using stereo inputs, while Wang *et al*. [39] do so with a single image. Li *et al*. [19, 23] design physically-based representations and rendering layers to estimate shape, SVBRDF, and spatially-varying per-pixel lighting from a single image. However, the aforementioned methods utilize convolutional neural networks, which feature a limited receptive field and lack an explicit attention mechanism to reason about long-range dependencies in image space, which can be crucial for estimating global properties of light transport and its interaction with material and shape.

Datasets for inverse rendering. Synthetic datasets [23, 30, 35] are commonly used to provide ground truth for most modalities, including scene geometry, material, and lighting, and suitable for training inverse rendering models in supervised fashion [19], while achieving good generalizability to real world datasets. Models trained on synthetic datasets are shown to further improve on real world images via finetuning with either weak supervision [4], full supervision on a subset of modalities [34], or with re-rendering losses [33]. In this work, we train our transformer models on the OpenRooms dataset [23] and obtain state-of-the-art results by finetuning on IIW [4] and NYUv2 [34] real world datasets.

Vision transformer. Convolutional neural networks (CNN) have long been the architecture of choice as building blocks for dense prediction with deep learning, as both backbone for feature extraction [5, 14] or as decoders [19, 27]. However, several drawbacks inherent to CNNs make them sub-optimal for tasks that require reasoning over long-range dependen-

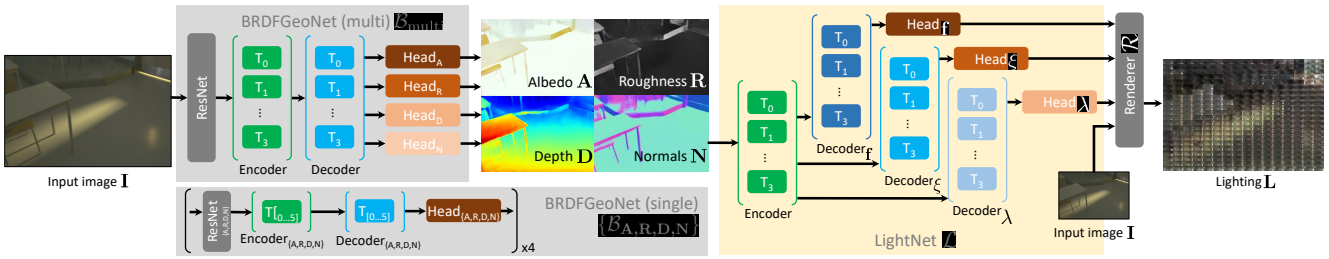

Figure 2. Overview of IRISformer. For **BRDFGeoNet**, we illustrate the multi-task setting in the upper gray block, while the single-task case (lower gray block) has 4 independent copies of DPT with different design of output heads.

cies over the image space, despite multiple measures that have been proposed to mitigate those issues, including dilated convolutions [5,42], skip connections [14,31] and self-attention [45]. The recently proposed Vision Transformer (ViT) [7] has enabled feature extraction with global attention over image space with elegant design and better interpretability while achieving superior performance compared to CNNs on multiple vision tasks. Several works [38,41,43] have extended ViT to dense prediction tasks, including DPT [28], Swin Transformer [25], etc. Additional efforts have been made to utilize transformers in a multi-task or multi-object setting, such as UniT [16] that follows an encoder-decoder design scheme and utilizes a task-specific query embedding to learn a unified decoding feature space for all tasks. In contrast to those works, we propose single-task and multi-task transformers for dense prediction tasks in inverse rendering, where material and lighting estimation using transformers have previously not been studied.

3. Method

Notation. Vectors are represented with a lower-case bold font (*e.g.* $\mathbf{x}$). Matrices are in upper-case bold (*e.g.* $\mathbf{X}$) while scalars are in regular font (*e.g.* x or X). Variables with hat, *e.g.* $\hat{\mathbf{X}}$, are the estimation of the corresponding entity $\mathbf{X}$. For denoting the l^{th} sample in a set (*e.g.* images, shapes), we use subscripts (*e.g.* $\mathbf{X}_l$). Uppercase calligraphic symbols (*e.g.* $\mathcal{X}$) denote functions.

3.1. Scene Representation and Loss Functions

Geometry and spatially-varying material. In IRISformer, we account for only the scene elements within the camera frustum. For an $h \times w$ image, we represent per-pixel geometry with a depth map $\mathbf{D} \in \mathbb{R}^{h \times w}$ and normals $\mathbf{N} \in \mathbb{R}^{h \times w \times 3}$. We represent material as a microfacet spatially-varying BRDF model (SVBRDF) [17], with albedo $\mathbf{A}$ and roughness $\mathbf{R}$ maps of sizes $h \times w \times 3$ and $h \times w$ respectively. Specifically, given a single RGB input image $\mathbf{I}$, we seek to learn a function **BRDFGeoNet** denoted as $\mathcal{B}$ to jointly estimate the above properties: $\{\hat{\mathbf{D}}, \hat{\mathbf{N}}, \hat{\mathbf{A}}, \hat{\mathbf{R}}\} = \mathcal{B}(\mathbf{I})$. Given ground truths, we use scale-invariant L2 loss [19,29] for albedo and depth (log space) and an L2 loss for roughness and normals.

Spatially-varying lighting. For lighting, we follow Li *et al.* [19,23] to adopt per-pixel image-space environment maps $\mathcal{G}$ of 16×32 pixels to represent the incident irradiance, parameterized by $K = 12$ Spherical Gaussian (SG) lobes $\{\xi_k, \lambda_k, \mathbf{f}_k\}_{k=1}^{K}$. Here $\xi_k \in \mathbb{S}^2$ is the center orientation on the unit sphere, $\mathbf{f}_k \in \mathbb{R}^3$ is the intensity and $\lambda_k \in \mathbb{R}$ is the bandwidth. Given each set of lighting parameters and an orientation outward from one spatial location η in 3D, we have

$$\mathcal{G}(\eta) = \sum_{k=1}^{K} \mathbf{f}_k e^{\lambda(1-\eta\xi)} : \mathbb{S}^2 \rightarrow \mathbb{R}^3. \quad (1)$$

For a collection of all $h \times w$ pixel locations and all 16×32 outgoing directions from each surface point in 3D, we arrive at a lighting map $\mathbf{L} \in \mathbb{R}^{h \times w \times 16 \times 32 \times 3}$. Then, using either a single image as the sole input or combining it with predictions of geometry and material, a **LightNet** denoted as $\mathcal{L}$ can be learned: $\hat{\mathbf{L}} = \mathcal{L}(\mathbf{I})$ or $\hat{\mathbf{L}} = \mathcal{L}(\mathbf{I}, \hat{\mathbf{D}}, \hat{\mathbf{N}}, \hat{\mathbf{A}}, \hat{\mathbf{R}})$. Given estimates of geometry, material and lighting, a per-pixel physically-based differentiable renderer **RenderNet** [19] denoted as $\mathcal{R}$ can re-render the input image: $\hat{\mathbf{I}} = \mathcal{R}(\hat{\mathbf{D}}, \hat{\mathbf{N}}, \hat{\mathbf{A}}, \hat{\mathbf{R}}, \hat{\mathbf{L}})$.

To supervise the training of lighting, assuming dense ground truth can be acquired from synthetic datasets, we impose a scale-invariant L2 reconstruction loss on the lighting map $\hat{\mathbf{L}}$ (in log space) and a scale-invariant L2 re-rendering loss on the re-rendered image $\hat{\mathbf{I}}$. The final loss L_{all} is a weighted combination of losses on all estimations:

$$L_{\text{all}} = \lambda_{\mathbf{A}} L_{\mathbf{A}} + \lambda_{\mathbf{R}} L_{\mathbf{R}} + \lambda_{\mathbf{D}} L_{\mathbf{D}} + \lambda_{\mathbf{N}} L_{\mathbf{N}} + \lambda_{\mathbf{L}} L_{\mathbf{L}} + \lambda_{\mathbf{I}} L_{\mathbf{I}}. \quad (2)$$

3.2. Dense Vision Transformer

Dense Vision Transformer (DPT) [28] is a generic architecture for dense prediction utilizing vision transformers [7] in place of CNNs as a backbone. An image $\mathbf{I}$ is first divided into an $h/p \times w/p$ grid of non-overlapping patches of size $p \times p$, and subsequently DPT tokenizes each patch into a vector of dimension D with a shadow CNN (DPT-large or DPT-base) or ResNet (DPT-hybrid). The result is a set of tokens $\mathbf{t}^0 = \{\mathbf{t}_1^0, \ldots, \mathbf{t}_{N_p}^0\}$ where $N_p = hw/p^2$, $\mathbf{t}_n^0 \in \mathbb{R}^D, n = [1, \ldots N_p]$. A cascade of M transformer layers then transforms the set of vectors with self-attention [37]

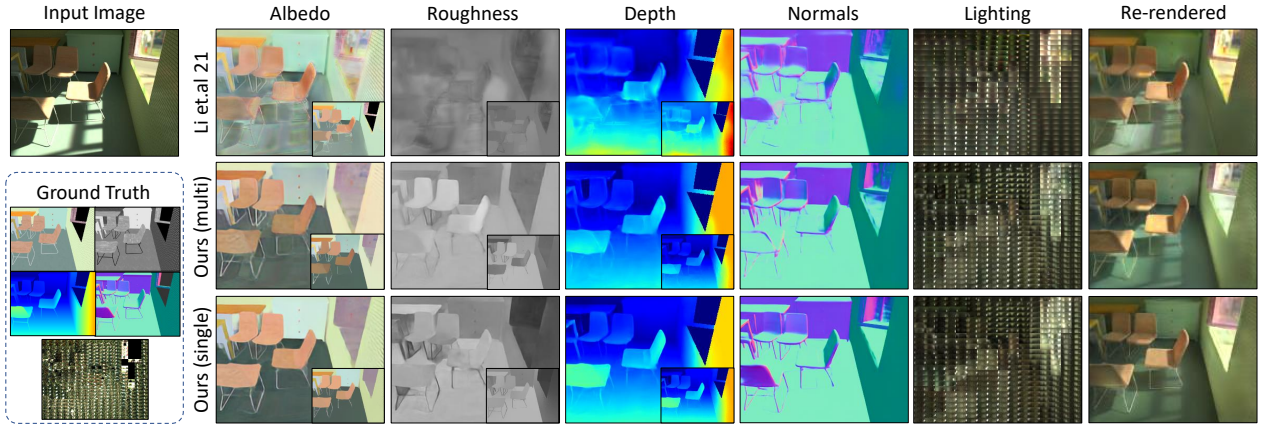

Figure 3. BRDF, geometry and lighting estimation on OpenRooms. Small insets (best viewed when enlarged in PDF version) are estimations processed with bilateral solvers (BS). More results can be found in the supplementary material.

into $\mathbf{t}^M$, and a re-assembling operation followed by a convolutional decoder transforms the tokens back to 2D space, resulting in a 2D dense feature map. A customized convolutional head is attached to yield the final prediction from the feature map, based on the specific prediction task.

3.3. Single-task Network Design

Due to the modular design of DPT and the variation of size and capacity among different DPT variants [28], we consider a few design choices for using DPT modules to build **BRDFGeoNet** and **LightNet** in both single-task and multi-task settings.

The full design of our pipeline can be found in Fig. 2. In single-task setting, we seek to maximize the performance of each task by using an independent DPT for each of depth, normal, albedo, roughness, and lighting. This effectively results in **BRDFGeoNet** of 4 DPTs $\{\mathcal{B}_A, \mathcal{B}_R, \mathcal{B}_D, \mathcal{B}_N\}$ to independent infer each of the modality: $\mathbf{D} = \mathcal{B}_D(\mathbf{I})$, $\mathbf{N} = \mathcal{B}_N(\mathbf{I})$, $\mathbf{A} = \mathcal{B}_A(\mathbf{I})$, $\mathbf{R} = \mathcal{B}_R(\mathbf{I})$. For each DPT, we follow the design of DPT-hybrid [28] by using $M = 6$ transformer layers for encoding and $M = 6$ for decoding. In our case, we use an input resolution of 256×320 and patch size $p = 16$. A ResNet-50 [14] acts as the patch embedding backbone. For output head design, we take output features from stage 1 and 2 of ResNet, as well as output from layer 3 and 6 from decoder, fuse the representations and use 4 convolutions layers with 2 bilinear interpolation layers to produce the final output. Readout token [28] is set to *ignore* and batch normalization (BN) is enabled in output heads. The only difference among the DPTs is the output layer in the head, depending on the sub-task. We use *tanh* activation for all heads to output albedo, roughness, normals and inverse depth. Details on the tensor sizes and head design are in the supplementary material.

For **LightNet** we have a similar encoder-decoder design. Three independent heads are required for estimating axis center, intensity, and bandwidth of K Spherical Gaussians for each pixel. However we found that sharing decoders for the three tasks is not optimal as the output spaces of these tasks are very disparate, thus forcing a unified decoder feature space destabilizes training. As a result, we use a shared encoder but independent decoders and output heads for **LightNet**. We use 4 transformer layers in both encoders and decoders in the multi-task setting so that the entire model can fit into one GPU for joint training. We use a similar output layer design as aforementioned except for not using *tanh* for ξ_k but use normalization to unit norm instead.

3.4. Multi-task Network Design

The single-network design requires 4 DPTs for material and geometry and one for lighting. As a result, the collective memory footprint is too large to fit the entire model into a regular GPU for training. An alternative option is to allow DPTs to have a unified feature space so that memory usage can be reduced. Also in some cases, a jointly learned feature can benefit from related tasks. Inspired by UniT [16] where input-domain-specific encoders and shared decoders are designed for a multi-modal input setting, we use a unified encoder and decoder for all tasks in **BRDFGeoNet** besides independent task-specific convolutional heads. We share decoders due to memory considerations while noting that further gains from independent decoders might be possible. The result is $\mathcal{B}_{multi}$ with 4 heads as shown in Fig. 2. The design for **LightNet** is the same as in a single-task setting.

3.5. Additional Components

Refinement using bilateral solver. We may optionally refine the geometry and material outputs with a bilateral solver (BS). Additional refinement leads to smoother outputs, which is preferable for some metrics like WHDR [4] for albedo. In comparison to previous CNN-based works, we observe that our transformer-based outputs are already quite accurate without the bilateral solver on all the tasks.

Cascade design. In prior works [19], a cascade design is

used to refine the predictions based on the rendering error. However, it leads to a two-fold increase in memory with small improvements, while we already achieve significantly improved results on all benchmarks with a single-stage network. As a result, we choose not to use cascaded refinement.

4. Evaluation

We demonstrate the capability of our DPT-based IRISformer to produce globally coherent estimations that outperform traditional CNN-based models on all modalities, due to its global attention that can better handle the inherent ambiguities of inverse rendering. This is especially notable for material and lighting prediction, in the presence of highlights, shadows, and interreflections. We include results on joint BRDF, geometry, and lighting prediction on OpenRooms, as well as sub-tasks on real world benchmarks. We additionally provide analysis on design choices and ablation study.

4.1. Datasets and Training

Given the success of synthetic inverse rendering datasets in providing photorealistic images and complete ground truth for all inverse rendering tasks, we use OpenRooms (OR) [23] dataset for supervised training. We use 6,684 scenes for training, 1,008 for testing, each rendered with multiple material and lighting configurations, for a total of 102,452 images for training, and 15,738 frames for testing.

We train **BRDFGeoNet** (in multi-task and single-task settings) on OR with Adam optimizer for 80 epochs at a learning rate of 1e-5 and batch size of 8 on 4 GPUs, starting with pretrained ResNet on ImageNet. Then we freeze **BRDFGeoNet** and train **LightNet** in the same setting. Additional training details and weights for losses are in the supplementary material.

After training on OpenRooms, we may finetune on real datasets where labels of all or a subset of the tasks are available. Specifically, we finetune on (a) IIW dataset [4] with relative labels of albedo; (b) NYUv2 [34] with ground truth depth and normals. We also demonstrate lighting estimation results with virtual object insertion on real world images from Garon *et al*. [12] where ground truth lighting is collected at selected locations with light probes.

4.2. BRDF and Geometry Estimation

Table 1 includes the performance of IRISformer as well as Li *et al*. [23] (which we trained and finetuned in the same setting as ours for all evaluations) for BRDF and geometry estimation, evaluated on our OpenRooms test split, both with a variant where the bilateral solver is applied to albedo, roughness, and depth. We observe better performance from both our multi-task and single-task models consistently on all tasks, and we visually demonstrate the comparison with a few samples in Fig. 3. As can be observed, our model excels with much cleaner and accurate material and geometry estimations, and better lighting estimations (especially in areas of highlights and shadows where we better disentangle lighting from albedo and correspondingly yields brighter/darker lighting).

Method	A↓	R↓	D↓	N↓	L↓	I↓	L+I↓
IRISformer (multi)	0.51	5.52	1.72	2.05	12.50	1.15	12.54
IRISformer (multi+BS)	0.51	5.50	1.71	2.05	12.47	1.15	12.58
IRISformer (single)	**0.43**	5.50	**1.42**	**1.89**	**12.04**	0.99	**12.14**
IRISformer (single+BS)	**0.43**	**5.48**	1.44	**1.89**	12.08	0.97	12.17
Ours (direct)	-	-	-	-	12.29	1.29	12.42
Li'21 [23]	0.52	6.31	2.20	2.61	18.63	**0.88**	18.72
Li'21+BS [23]	0.48	6.30	1.91	2.61	18.61	**0.88**	18.70

Table 1. Errors of BRDF, geometry and lighting with a base of 10^{-2} on OpenRooms [23]. Lower is better. For lighting estimation, **L** is the lighting reconstruction error, **I** is the rendering error and **L+I** is the combined lighting loss for which **LightNet** is trained.

4.3. Lighting Estimation

In Table 1, we report lighting estimation errors of different versions of IRISformer, including both multi-task and single-task models, and variants of these two models with BS. We also include another version of our lighting estimation model (denoted as *Ours (direct)*), where the first stage of BRDF and geometry estimation is removed, and only the image is directly used as input for lighting prediction. We observe the best lighting prediction from the single models due to their larger capacity. For the direct model, it is able to estimate reasonable lighting directly from image input, highlighting the power of the transformer architecture, but is not suitable for downstream tasks (*e.g*. object insertion, material editing) where a complete scene decomposition is required.

4.4. Comparison on Sub-tasks

Intrinsic decomposition. To evaluate IRISformer on the task of intrinsic decomposition (albedo-only) on real world images, we finetune our model on IIW [4] using relative labels of albedo between labeled pairs of pixels. Results are summarized in Table 2 and Fig. 5. We observe that our single-task model performs better than the multi-task version due to its larger capacity and independent feature space. More importantly, both the multi-task and single-task models achieve new state-of-the-art on IIW, outperforming all prior methods. Note that the shift in tones is due to the weak supervision from relative loss.

Geometry estimation. For depth and normal prediction, we follow the training and evaluation settings of Li *et al*. [23], and report results in Table 3 and Fig. 5. We choose to compare with similar methods in a multi-task inverse rendering setting, instead of dedicated and more complex methods that maximize geometry prediction performance in the wild like DPT [28] or MiDaS [29]. As can be observed, we achieve improved results compared to all previous works listed.

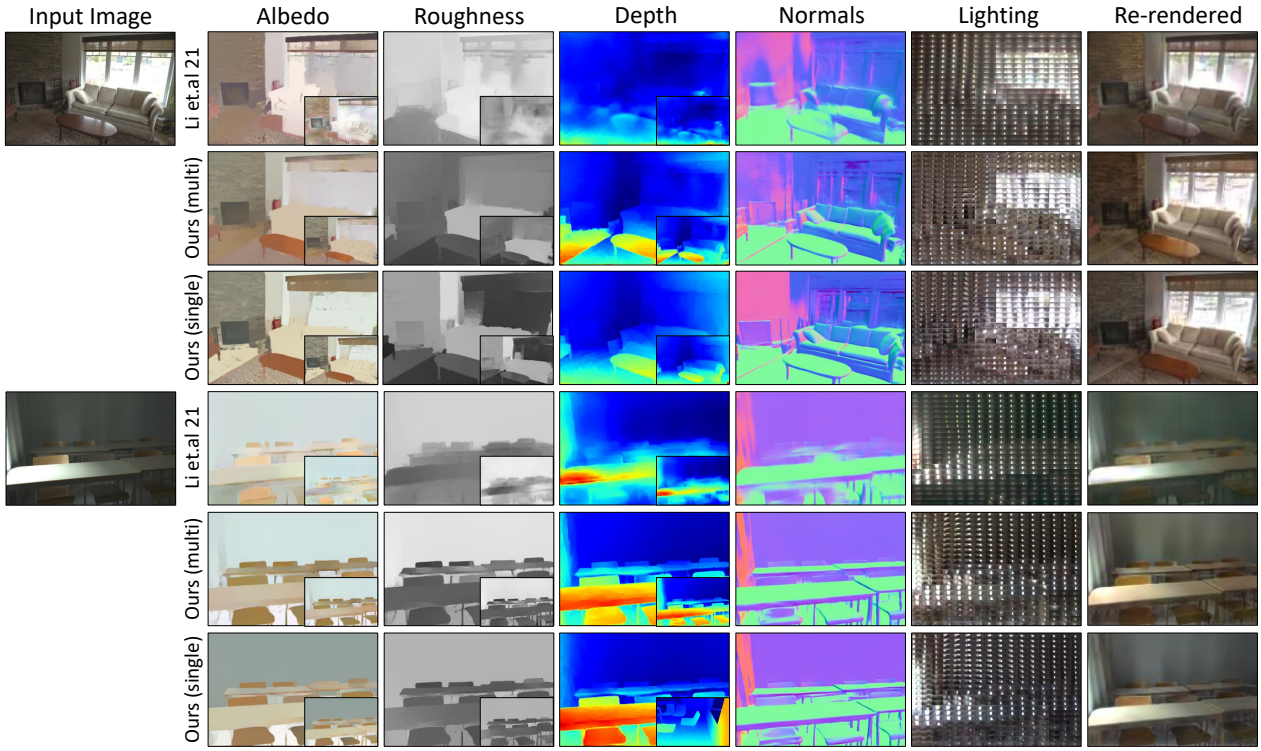

Figure 4. BRDF, geometry estimation, per-pixel lighting and re-rendering results on Garon *et al*. [12] (after BS). Insets are results before BS. Our material and geometry results are cleaner even without BS. Also less artifacts can be observed in our re-rendered images (see bright area on the table in sample 2) compared to Li *et al*. [23] due to better estimation on all modalities.

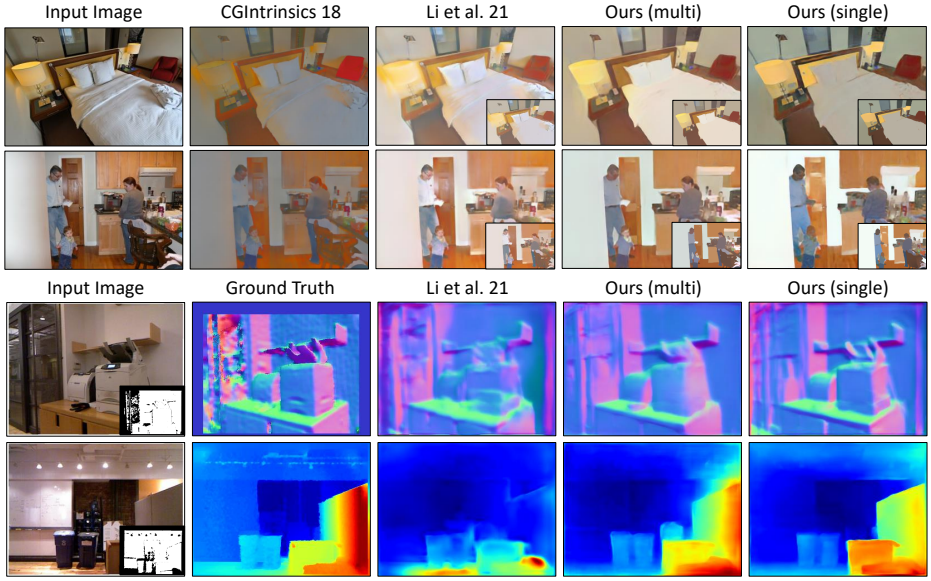

Figure 5. (Top) Intrinsic decomposition results on IIW [4] (before BS). The insets are results after BS. Our results are better in regions with complex geometry and lighting (see the bedding in sample 1 and clothing in sample 2). (Bottom) Geometry estimation results on NYUv2 [34] (all without BS). Ours are less prone to artifacts (see printer surface in sample 1 and geometry of the trash bins in sample 2). Please refer to the supplementary material for more results.

Intermediate results on real world images. In Fig. 4 we test IRISformer on real world images from Garon *et al*. [12], and we demonstrate that IRISformer generalize well to real world images and outperform previous art in every task. In general, our results are more spatially consistent and have fewer artifacts (most noticeable from the re-rendered images which are the result of all estimations). In Fig. 6, we compare with the recent method of Wang *et al*. [39] on their reported samples, where we arrive at much improved results.

Lighting estimation on real images. With intermediate estimations including material, geometry, and the final per-pixel lighting, we demonstrate applications of IRISformer in downstream applications including virtual object insertion and material editing. For object insertion, we compare with

Method	Finetuning Datasets	WHDR↓
IRISformer (multi)	OR+IIW	13.1
IRISformer (single)	OR+IIW	**12.0**
Li'21 [23]	OR+IIW	16.4
Li'20 [19]	CGM+IIW	15.9
NIR'19 [33]	CGP+IIW	16.8
CGIntrinsics'18 [20]	CGI+IIW	17.5

Table 2. Intrinsic decomposition on IIW [4]. Lower is better.

Method	Mean(°)↓	Med.(°)↓	Depth↓
IRISformer (multi)	23.5	16.3	0.162
IRISformer (single)	**20.2**	**13.4**	**0.132**
Li'21 [23]	25.3	18.0	0.171
Li'20 [19]	24.1	17.3	0.184
NIR'19 [33]	21.1	16.9	-
Zhang'17 [47]	21.7	14.8	-

Table 3. Normal (mean and median) and depth (mean on inverse depth) prediction results on NYUv2 [34]. Lower is better.

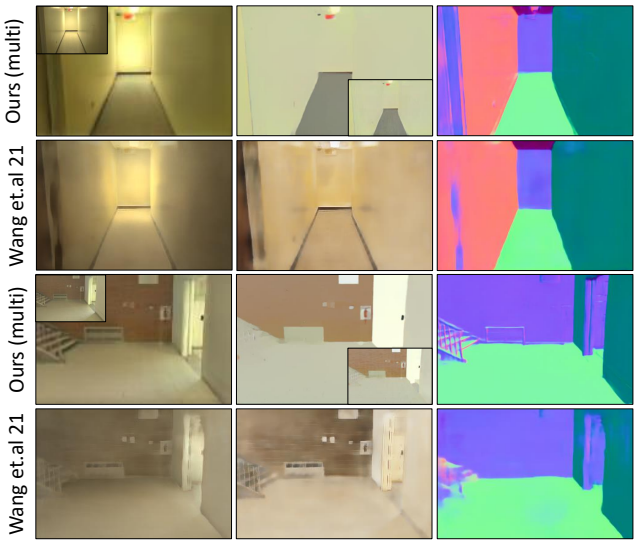

Figure 6. Comparison between our results and Wang *et al*. [39] on albedo, normals and re-rendering. On challenging inputs, we achieve smoother albedo with less artifacts and richer details with more consistent normals. Upper-left insets are input images, and lower-right insets are before BS).

Gardner'17 [11]	Garon'19 [12]	Li'21 [23]	Ground Truth
0.24	0.30	0.47	0.58

Table 4. A user study on object insertion, where we compare IRISformer with each of the previous work or ground truth and report the percentage of feedbacks where other method is considered to be more photorealistic than ours.

prior works in Fig. 9. We produce more realistic lighting for inserted objects which better match the surroundings on lighting intensity, direction and relative brightness of inserted objects in highlights and shadowed areas.

To quantitatively evaluate insertion results, we conduct a user study against other methods in Tab. 4. We outperform all previous methods, only being inferior to ground truth. We also perform material editing in Fig. 7 where we replace the material of a planar surface and re-render the region, to showcase that IRISformer captures the directional lighting effect of the area so that the replaced material will be properly shadowed. A complete list of results and comparisons is in the supplementary material.

	single-6	single-4	multi	Li'20 [23]
Model Size (MB)	7,305	6,256	1,539	795
Inference (ms)	141.9	125.9	91.9	45.2
A+R+D+N	6.00	6.08	6.44	7.65
L+I	12.14	12.85	12.54	18.72

Table 5. Analysis on multiple design choices: IRISformer (single-task with 6 or 4 layers in **BRDFGeoNet**, multi-task with 4 layers), and CNN-based architecture from Li *et al*. [23] on OR [23].

Figure 7. Material editing example where we replace the material of part of the wall with wood. Note that the shadow from outside lighting is recreated on the replaced material, demonstrating our accurate spatially-varying lighting estimation.

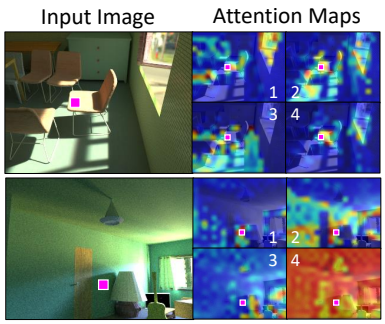

Figure 8. Attention maps learned by the single-task model for albedo. Each heatmap is the attention weights (affinity) of the patch (denoted by pink square) w.r.t. all other patches, of one head from subsequent transformer layers.

4.5. Ablation study

Comparison of design choices. In Table 5 we compare various metrics of all models, including model size, inference time for one sample on a Titan RTX 2080Ti GPU, test losses on OR for material and geometry combined, and lighting losses combined. Study on other minor choices can be found in the supplementary material.

Attention from Transformers. To provide additional insight into the attention that is learned, we include in Fig. 8 two samples where attention maps corresponding to one patch location from different layers are visualized. In the first sample we show a patch on the lit-up chair seat, subsequently attending to (1) chairs and window, (2) highlighted regions over the image, (3) entire floor, (4) the chair itself. For the second sample, the chosen patch is in the shadow on

Figure 9. Virtual object insertion results. Lighting estimation from Barron *et al*. [2] lacks high frequency, while Gardner *et al*. [11] predicts a global environment map instead of spatially-varying lighting at each location. Compared to the most recent Li *et al*. [23], we better decouple lighting and appearance to better recover highlighted or shadowed areas (see center object in sample 1 and 2, right object in sample 4 and 5). Also our lighting is more spatially consistent w.r.t. the light source (see shadow directions in objects of sample 1 and 3).

the wall, and it attends to (1) neighbouring shadowed areas of the wall, (2) the entire wall, (3) potential light sources and occluders, (4) ambient environment. Throughout the cascaded transformer layers, IRISformer learns to attend to large regions and distant interactions to improve its prediction in the presence of complex light transport effects.

5. Discussion

Limitation and potential negative impact. IRISformer only infers per-pixel lighting on the scene surface, so applications like inserting objects in the air are not feasible. Future work may also explore choices beyond the current multi-task design, possibly by leveraging the complementary nature of various tasks. Potential negative impacts include Deepfake [40], where our method can be used to recreate an indoor scene with a photorealistically modified appearance.

Conclusion. We have proposed an inverse rendering framework that estimates material, geometry, and per-pixel lighting given an unconstrained indoor image using a transformer-based model. Our results demonstrate that the model can produce significantly better results especially on material and lighting, which require long-range reasoning for diambiguation. Additionally, our approach enables different design choices with single or multi-task settings. Downstream applications including object insertion and material editing on real world images demonstrate the strength of our model to better handle challenging lighting conditions and produce highly photorealistic results. We also provide analysis into design choices and the attention maps learned by our model.

Acknowledgments: We thank NSF CAREER 1751365, NSF IIS 2110409 and NSF CHASE-CI, generous support by Qualcomm, as well as gifts from Adobe and a Google Research Award.

References

[1] Miika Aittala, Timo Aila, and Jaakko Lehtinen. Reflectance modeling by neural texture synthesis. *ACM Trans. Graphics*, 35(4):65, 2016. 2

[2] Jonathan T Barron and Jitendra Malik. Intrinsic scene properties from a single rgb-d image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 17–24, 2013. 1, 2, 8

[3] Jonathan T Barron and Jitendra Malik. Shape, illumination, and reflectance from shading. *IEEE transactions on pattern analysis and machine intelligence*, 37(8):1670–1687, 2014. 1, 2

[4] Sean Bell, Kavita Bala, and Noah Snavely. Intrinsic images in the wild. *ACM Transactions on Graphics (TOG)*, 33(4):1–12, 2014. 1, 2, 4, 5, 6, 7

[5] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2, 3

[6] Valentin Deschaintre, Miika Aittala, Fredo Durand, George Drettakis, and Adrien Bousseau. Single-image svbrdf capture with a rendering-aware deep network. *ACM Trans. Graphics*, 37(4):128, 2018. 2

[7] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 3

[8] David Eigen and Rob Fergus. Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In *Proceedings of the IEEE international conference on computer vision*, pages 2650–2658, 2015. 1

[9] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-François Lalonde. Deep parametric indoor lighting estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7175–7183, 2019. 1

[10] Marc-André Gardner, Yannick Hold-Geoffroy, Kalyan Sunkavalli, Christian Gagné, and Jean-François Lalonde. Deep parametric indoor lighting estimation. In *Proc. ICCV*, 2019. 2

[11] 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 Transactions on Graphics (TOG)*, 36(6):1–14, 2017. 1, 2, 7, 8

[12] Mathieu Garon, Kalyan Sunkavalli, Sunil Hadap, Nathan Carr, and Jean-François Lalonde. Fast spatially-varying indoor lighting estimation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6908–6917, 2019. 5, 6, 7

[13] Roger Grosse, Micah K Johnson, Edward H Adelson, and William T Freeman. Ground truth dataset and baseline evaluations for intrinsic image algorithms. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2335–2342. IEEE, 2009. 2

[14] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 3, 4

[15] Berthold KP Horn and Michael J Brooks. *Shape from shading*. MIT press, 1989. 2

[16] Ronghang Hu and Amanpreet Singh. Unit: Multimodal multitask learning with a unified transformer. *arXiv preprint arXiv:2102.10772*, 2021. 3, 4

[17] Brian Karis and Epic Games. Real shading in unreal engine 4. *Proc. Physically Based Shading Theory Practice*, 4(3), 2013. 3

[18] Chloe LeGendre, Wan-Chun Ma, Graham Fyffe, John Flynn, Laurent Charbonnel, Jay Busch, and Paul Debevec. Deeplight: Learning illumination for unconstrained mobile mixed reality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5918–5928, 2019. 1

[19] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2475–2484, 2020. 1, 2, 3, 4, 7

[20] Zhengqi Li and Noah Snavely. Cgintrinsics: Better intrinsic image decomposition through physically-based rendering. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 371–387, 2018. 1, 2, 7

[21] Zhengqin Li, Kalyan Sunkavalli, and Manmohan Chandraker. Materials for masses: SVBRDF acquisition with a single mobile phone image. In *ECCV*, 2018. 2

[22] Zhengqin Li, Zexiang Xu, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Learning to reconstruct shape and spatially-varying reflectance from a single image. In *SIGGRAPH Asia*, page 269. ACM, 2018. 2

[23] Zhengqin Li, Ting-Wei Yu, Shen Sang, Sarah Wang, Meng Song, Yuhan Liu, Yu-Ying Yeh, Rui Zhu, Nitesh Gundavarapu, Jia Shi, Sai Bi, Zexiang Xu, Hong-Xing Yu, Kalyan Sunkavalli, Miloš Hašan, Ravi Ramamoorthi, and Manmohan Chandraker. OpenRooms: An end-to-end open framework for photorealistic indoor scene datasets. In *CVPR*, 2021. 2, 3, 5, 6, 7, 8

[24] Chen Liu, Kihwan Kim, Jinwei Gu, Yasutaka Furukawa, and Jan Kautz. Planercnn: 3d plane detection and reconstruction from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4450–4459, 2019. 1

[25] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *arXiv preprint arXiv:2103.14030*, 2021. 2, 3

[26] Steve Marschner. Inverse rendering for computer graphics. 1998. 2

[27] Alejandro Newell, Kaiyu Yang, and Jia Deng. Stacked hourglass networks for human pose estimation. In *European conference on computer vision*, pages 483–499. Springer, 2016. 2

[28] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12179–12188, 2021. 2, 3, 4, 5

[29] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *arXiv preprint arXiv:1907.01341*, 2019. 3, 5

[30] Mike Roberts and Nathan Paczan. Hypersim: A photorealistic synthetic dataset for holistic indoor scene understanding. *arXiv preprint arXiv:2011.02523*, 2020. 2

[31] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 3

[32] Shen Sang and Manmohan Chandraker. Single-shot neural relighting and svbrdf estimation. In *ECCV*, 2020. 2

[33] Soumyadip Sengupta, Jinwei Gu, Kihwan Kim, Guilin Liu, David W Jacobs, and Jan Kautz. Neural inverse rendering of an indoor scene from a single image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8598–8607, 2019. 1, 2, 7

[34] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *European conference on computer vision*, pages 746–760. Springer, 2012. 2, 5, 6, 7

[35] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1746–1754, 2017. 2

[36] Pratul P Srinivasan, Ben Mildenhall, Matthew Tancik, Jonathan T Barron, Richard Tucker, and Noah Snavely. Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8080–8089, 2020. 1, 2

[37] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 3

[38] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without cowhat nvolutions. *arXiv preprint arXiv:2102.12122*, 2021. 2, 3

[39] Zian Wang, Jonah Philion, Sanja Fidler, and Jan Kautz. Learning indoor inverse rendering with 3d spatially-varying lighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12538–12547, 2021. 1, 2, 6, 7

[40] Mika Westerlund. The emergence of deepfake technology: A review. *Technology Innovation Management Review*, 9(11), 2019. 8

[41] Chenfeng Xu, Bohan Zhai, Bichen Wu, Tian Li, Wei Zhan, Peter Vajda, Kurt Keutzer, and Masayoshi Tomizuka. You only group once: Efficient point-cloud processing with token representation and relation inference module. *arXiv preprint arXiv:2103.09975*, 2021. 3

[42] Fisher Yu and Vladlen Koltun. Multi-scale context aggregation by dilated convolutions. *arXiv preprint arXiv:1511.07122*, 2015. 3

[43] Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zihang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. Tokens-to-token vit: Training vision transformers from scratch on imagenet. *arXiv preprint arXiv:2101.11986*, 2021. 2, 3

[44] Fangneng Zhan, Yingchen Yu, Rongliang Wu, Changgong Zhang, Shijian Lu, Ling Shao, Feiying Ma, and Xuansong Xie. Gmlight: Lighting estimation via geometric distribution approximation. *arXiv preprint arXiv:2102.10244*, 2021. 2

[45] Han Zhang, Ian Goodfellow, Dimitris Metaxas, and Augustus Odena. Self-attention generative adversarial networks. In *International conference on machine learning*, pages 7354–7363. PMLR, 2019. 3

[46] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah. Shape-from-shading: a survey. *IEEE transactions on pattern analysis and machine intelligence*, 21(8):690–706, 1999. 2

[47] Yinda Zhang, Shuran Song, Ersin Yumer, Manolis Savva, Joon-Young Lee, Hailin Jin, and Thomas Funkhouser. Physically-based rendering for indoor scene understanding using convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5287–5295, 2017. 7