STAG: Smart Tools and Applications in Graphics (2024)
A. Caputo, V. Garro, U. Castellani, and A. Giachetti (Editors)

VISPI: Virtual Staging Pipeline for Single Indoor Panoramic Images

Uzair Shah[1], Sara Jashari[1], Muhammad Tukur[1], Giovanni Pintore[2,3], Enrico Gobbetti[2,3], Jens Schneider[1], Marco Agus[1]

[1]College of Science and Engineering, Hamad Bin Khalifa University, Qatar;
[2]National Research Center in HPC, Big Data, and QC, Italy;
[3]CRS4, Italy;

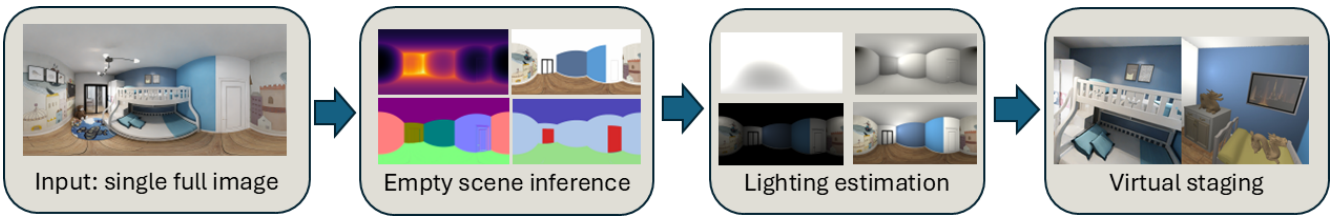

Figure 1: ***VISPI:*** *our novel virtual staging pipeline takes as input a single omnidirectional image representing an indoor environment, performs deep-learning-based inference of multiple signals related to the scene without clutter, and estimates lighting through Spherical Gaussian parameterization in a way to allow placement of virtual objects inside the scene with plausible lighting conditions.*

Abstract

Taking a 360° image is the quickest and most cost-effective way to capture the entire environment around the viewer in a form that can be directly exploited for creating immersive content [PBAG23]. In this work, we introduce novel solutions for the virtual staging of indoor environments, supporting automatic emptying, object insertion, and relighting. Our solution, dubbed VISPI (Virtual Staging Pipeline for Single Indoor Panoramic Images), integrates data-driven processing components, that take advantage of the analysis of knowledge learned from massive data collections, within a real-time rendering and editing system, allowing for interactive restaging of indoor scenes. Key components of VISPI include: i) a holistic architecture based on a multi-task vision transformer for extracting geometry, semantic, and material information from a single panoramic image, ii) a lighting model based on spherical Gaussians, iii) a method for lighting estimation from the geometric, semantic, and material signals, and iv) a real-time editing and rendering component. The proposed framework provides an interactive and user-friendly solution for creating immersive visualizations of indoor spaces. We present a preliminary assessment of VISPI using a synthetic dataset – Structured3D – and demonstrate its application in creating restaged indoor scenes.

CCS Concepts

• ***Computing methodologies*** → ***Computer graphics; Computer vision; Shape inference; Neural networks;***

1. Introduction

Virtual staging is becoming an increasingly popular tool in industries such as real estate, interior design, and architecture, where visualization of alternative designs plays a crucial role in decision-making processes [ZCB*22]. Traditional virtual staging methods, often relying on flat 2D images, have several limitations. These methods usually require multiple images to cover a space adequately, leading to a fragmented and sometimes disjointed representation of the environment. Moreover, the lack of a comprehensive, immersive view can hinder the user's ability to fully experience and evaluate a space.

To overcome these limitations, omnidirectional imaging, a technique that captures most of the visual information about a scene in a single shot, has become increasingly popular for acquiring environments in the Architecture, Engineering, and Construction (AEC) domain, particularly for indoor scenes [PGGS16, SLK*23].

Concurrently, several data-driven solutions have been developed for a range of processing operations that are now routinely performed in the AEC domain, such as the extraction of metric information about room layouts [PAG20], 3D geometry in the form of meshes [PAAG21] or point clouds, and semantic information to enable editing operations [SLL*22].

More recently, researchers have begun exploring diminished reality solutions that can be applied for restaging applications [JSN24]. These solutions effectively empty the interior of an environment, making it ready to be filled with different 3D as-

DOI: 10.2312/stag.20241334

sets to create new environments [PAAG22]. However, to achieve realism and effectiveness, it is essential to recover also information about the lighting conditions. Although recent advancements in inverse rendering techniques have been proposed to address this need [GDHG*24], they primarily focus on perspective images and have complexity requirements that limit their integration into real-time editing pipelines.

To address these gaps and overcome these challenges, we propose a novel processing framework named VISPI (Virtual Staging Pipeline for Single Indoor Panoramic Images). VISPI is designed to enhance virtual staging by focusing on single omnidirectional images. It integrates data-driven processing components within a real-time rendering and editing system, enabling the interactive restaging of indoor scenes represented by single panoramic images. The framework includes the following novel components (see Fig. 1):

- An architecture based on a vision transformer for multi-task extraction of geometry, semantic, and material information from single equirectangular RGB images of fully cluttered scenes (Sec. 4). This architecture is exploited to create uncluttered environments ready for virtual staging.
- A parametric lighting model based on spherical Gaussians computed from the geometry and material information related to the indoor scene. This model is used for two tasks: estimating the light parameters through the least square method from inferred signals, and illuminating virtual objects through a real-time shader (Sec. 5).
- A real-time editing and rendering component that integrates all precomputed scene information, enabling virtual object placement inside the empty environment.

The proposed framework offers a fast interactive solution for creating immersive visualizations of indoor spaces. Our virtual staging pipeline has transformative potential across several application domains: real estate, furniture retail, interior design, the construction industry, as well as the creation of immersive environments for the Metaverse for supporting remote collaboration, and immersive training [TSH*24, CA24]. We report on a preliminary assessment of the pipeline on a synthetic public domain dataset (Structured3D [ZZL*20]) and we describe examples of usage of the pipeline for the creation of restaged indoor scenes.

2. Related work

Our work deals with data-driven processing solutions single panoramic images representing indoor environments and inverse rendering methods targeting virtual staging applications. For space reasons, we won't provide here an extensive discussion of the important literature corpus in this field, but we will instead discuss the most recent methods most closely related to our framework. For a comprehensive literature review related to these topics, we refer the interested readers to the recent surveys about omnidirectional visual computing [dJ23], indoor reconstruction [PMG*20, WL24], 3D scene understanding from panoramic imaging [dSPMLJ22], and deep learning for lighting estimation [EGH21].

Indoor panoramic processing Inferring geometric and physical signals from omnidirectional images presents a challenging problem that has garnered significant attention from the computer vision community in recent years. Generally, to effectively manage the spherical distortions caused by equirectangular projection, various approaches have been developed that incorporate reprojection in conjunction with Convolutional Neural Networks. More specifically, several methods have been proposed to tackle individual dense estimation tasks. For geometry estimation, methods like UniFuse [JSZ*21] and Bifuse [WYS*20] significantly reduce distortion by combining features extracted from both equirectangular projections and cube maps at different stages within encoder-decoder architectures. Meanwhile, M3PT [YLW*22] applies random masking to process both panoramas and sky-box depth patches simultaneously, targeting panoramic depth completion. Other techniques involve slicing the panoramic image vertically, based on the assumption that vertical lines remain undistorted in equirectangular projections when the acquisition is mostly gravity-aligned [PAA*21, PAAG21]. Additionally, some methods employ tangent projection to extract multiple undistorted patches that are processed using standard approaches for perspective images, as demonstrated by 360MonoDepth [RAYR22] and OmniFusion [LGY*22]. Recently, vision transformers have been explored for these tasks. PanoFormer [SLL*22] uses tangent projection (TP) to reduce the inherent distortion in omnidirectional images, treating the TP patches as tokens within a vision transformer architecture. HRDFuse [ACC*23] proposes a hybrid CNN-transformer architecture that integrates comprehensive contextual information from the original equirectangular projection with regional structural information extracted via tangent projection. Similarly, PanelNet [YHJ*23] represents the equirectangular projection (ERP) as sequential vertical panels with corresponding panel geometry, using a transformer to aggregate local information within a panel along with the panel-wise global context. Lastly, EGformer [YSL*23] focused on extracting equirectangular geometry-aware local attention with a large receptive field, by using the geometry as the bias for the local attention instead of trying to reduce the distortion generated by equirectangular projection. Another class of methods try to get additional information enabling editing operations on indoor environments, like the extraction of 3D layouts [PAAG21], or full scene composition [DFB*24] in form of 3D layouts together with the oriented bounding boxes of the objects contained inside the scene. Similarly, Pintore et al. [PAAG22] proposed a model for diminished reality able to automatically infer the depth signal of the scene without clutter. For what concerns Diminished Reality, very recently Gsaxner et al. [GMS*24] proposed a structure-aware generative network able to perform RGB-D inpainting in real time without artifacts, while Liu et al [LZS*23] proposed an impainting method for indoor panoramic images based on fast Fourier convolution. Finally, MultiPanoWise [STA*24] proposed a vision transformer for branched multi-task inference by introducing a hybrid Pareto-optimal hybrid loss scalarization strategy for improving the inference of multiple signals. In this work, we extend the latter architecture and take inspiration from the instant emptying method proposed by Pintore et al. [PAAG22]. The proposed architecture is a multi-task vision transformer able to infer multiple signals related to an empty scene from an RGB picture of a cluttered scene. Specifically, our model can extract concurrently depth, normals, semantic information, and reflectance information.

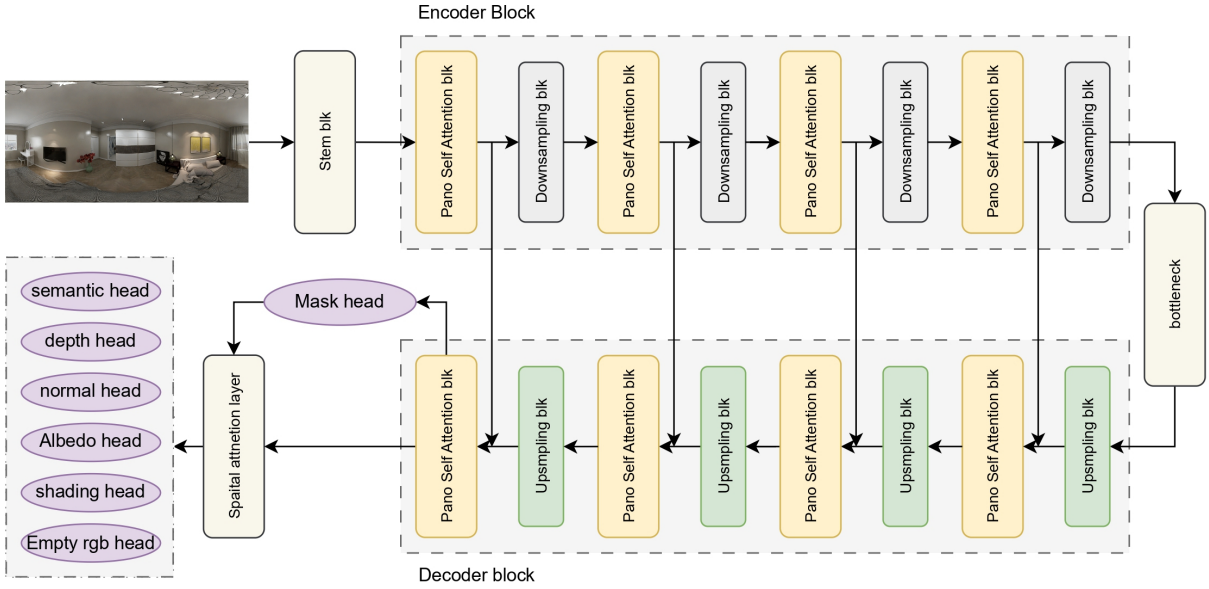

Figure 2: ***Proposed Architecture****: Our model builds upon the MultiPanoWise framework [SSP*24], with the addition of a new binary mask head. We compute spatial attention between the multi-task features and the binary mask to effectively filter out furniture and clutter. The refined multi-task features are then passed to multiple heads, generating outputs such as semantic segmentation, depth, and other task-specific signals.*

Inverse rendering and virtual staging Inverse rendering is the process of inferring scene properties from images in a way to enable editing and rendering applications [LWH*22], and it is a challenging problem. The task is ill-posed, as many different scene configurations can produce the same image [LLM21]. Over the last years, various solutions have been proposed to decompose scenes, most of them based on data-driven architectures, and falling into two main categories for indoor scenes. The first category considers parametric representations like spherical harmonics [GSH*19] and spherical Gaussians [GHGS*19, LSR*20, LYO*23], while the second category considers prediction models for environment maps [WYLL22, LMF*19]. Both methods have advantages and drawbacks: the parametric methods are simpler to implement but they often fail to capture correctly the lighting conditions, while the environment map prediction gives accurate results on perspective images but the underlying models are very complex to setup and train. A potential application of inverse rendering methods is scene editing, and especially virtual staging operations: recently, Zhi et al. [ZCB*22] developed a semantically supervised appearance decomposition architecture for performing virtual object insertion, while Ji et al. [JSN24] developed an inverse rendering approach using the High Dynamic Range (HDR) technique to capture an indoor panorama and its paired outdoor hemispherical photograph, for scene relighting and editing under natural illumination. In our pipeline, we consider a parametric model based on Spherical Gaussians, taking as input the inferred material, semantic, and geometry properties of indoor panoramic scenes, and providing as output a set of light parameters that can be used for virtual shading operations. We trained our inference and parametric model on synthetic data [ZZL*20].

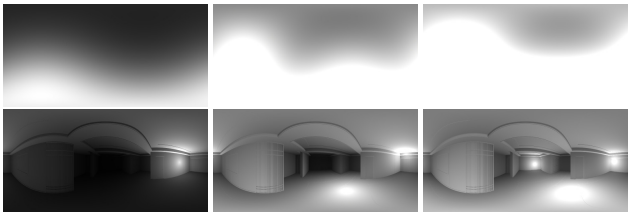

Figure 3: ***Evolution of light estimation with incremental Spherical Gaussian lights****. On the top the estimated light maps starting from one Spherical Gaussian light up to three. On the bottom are the approximated diffuse maps.*

3. Pipeline overview

The processing pipeline for virtual staging is depicted in Fig. 1. It is mainly composed by two subsequent tasks:

- **Empty scene inference**: We developed an end-to-end holistic architecture for multi-task learning, specifically designed for single panoramic images representing cluttered indoor environments. Our framework is based on the MultiPanoWise [SSP*24] architecture and is capable of concurrently inferring multiple signals relevant to tasks such as virtual staging or inverse rendering. These signals include geometric information in the form of depth and surface normals, semantic segmentation masks, and intrinsic decomposition signals like reflectance and shading.
  The core of our framework is an encoder-decoder structure that enriches multi-task features. These enriched features are passed through various task-specific heads, each responsible for inferring a particular signal. Before generating these signals, the multi-task feature is fed into a separate mask decoder that generates a binary mask, representing the probability of object presence. This binary mask assigns high probabilities to key archi-

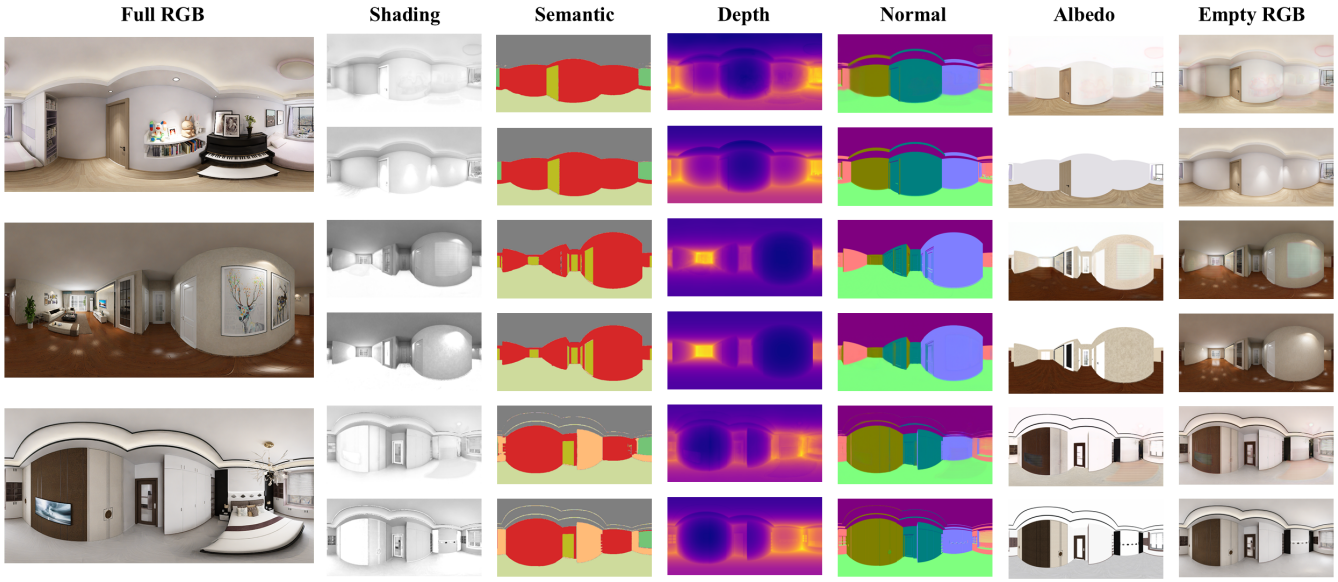

Figure 4: ***Qualitative Example of Model Predictions vs Ground Truth:*** *This figure illustrates the comparison between predicted and ground truth outputs across three scenes, with each row corresponding to one example. The columns, from left to right, show: (1) Full RGB Image, (2) Shading, (3) Semantic Segmentation, (4) Depth Map, (5) Normal Map, (6) Albedo, and (7) Empty RGB Scene. For each row, the top image represents the model's prediction, while the bottom image is the ground truth.*

tectural elements like walls, ceilings, floors, windows, and doors while assigning lower probabilities to other objects.
Next, we calculate spatial attention between the multi-task features and the binary mask to effectively mask out furniture and other objects that need removal. The refined multi-task features are then passed to their respective heads to generate various outputs such as empty RGB images, shading, albedo, normals, semantic segmentation, and depth maps. This process enables efficient processing of cluttered indoor scenes by extracting detailed and task-specific information.

- **Lighting estimation**: We developed a fast and interactive lighting estimation method for indoor omnidirectional images using Spherical Gaussian (SG) models [XSD*13]. The approach involves optimizing lighting parameters to minimize residuals between the ground truth and approximated images, enabling realistic shading and virtual placement of 3D objects with consistent lighting. The lighting is modeled using SG functions, which efficiently approximate specular and diffuse components. The algorithm iteratively adds Gaussian lights, optimizing the Gaussian light source parameters such as color and position through least-squares minimization. This method is also implemented as a GLSL shader in an OpenGL system for panoramic image rendering.

In the following we detail the components of the proposed pipeline.

4. Empty scene inference

Our framework is based on MultiPanoWise [SSP*24], a multi-task learning model designed for processing panoramic images as presented in Fig. 2. Building on the foundation of the PanoFormer transformer [SLL*22], this architecture enhances local feature extraction through pixel-level patch division, integrates relative position embeddings for better positional awareness, and employs panoramic self-attention to capture key structures in panoramic images. The model architecture is depicted in Fig. 2 and follows an encoder-decoder structure with four hierarchical stages, each consisting of position embeddings, Panoramic Self-Attention (PST) blocks, and convolution layers. The decoder is equipped with multiple heads that generate various outputs, including semantic segmentation, intrinsic decomposition signals (reflectance and shading), surface normals, and depth maps. To extend the model's capabilities, we introduced a new head for generating an empty RGB signal, which removes objects from the scene. Initially, this model served as our baseline, which we progressively refined to handle both object removal and the generation of multiple signals concurrently. Drawing inspiration from the Instant Empty framework [PAAG22], we adopted a two-stage approach. In the first stage, a binary mask is generated, marking cluttered pixels with 1 and non-cluttered pixels with 0. This binary mask is then combined with the original panoramic image and passed to the main model, which removes furniture and other objects while simultaneously generating empty RGB and depth outputs. By incorporating this two-stage approach—using a CNN-based U-Net architecture to generate the binary mask in the first stage—we observed significant performance improvements over the baseline model. However, challenges in generating precise signals remained. To address this, we added a new head within the main model to predict the binary mask directly. This additional head serves as a supervision mechanism, allowing the model to better identify objects for removal without relying on a separate model, as required in the Instant Empty approach. This modification reduced the number of

parameters and further improved the model's performance. Furthermore, we removed the context adjustment module from the original PanoWise architecture, which was originally intended to mitigate data loss and correct distortions, particularly around object edges. The module fused low-level features from the input stem with raw semantic, albedo, and shading outputs to fill gaps and correct distortions. However, our qualitative results revealed unwanted geometric structures and shadows in the final outputs, likely caused by the blending of low-level features with raw outputs. By removing this module, we achieved cleaner outputs and enhanced overall model performance. Finally, we compute spatial attention between the multi-task features and the binary mask to effectively filter out furniture and other objects marked for removal. The refined multi-task features are then passed to their respective heads to generate various outputs such as empty RGB images, shading, albedo, surface normals, semantic segmentation, and depth maps. This process enables efficient processing of cluttered indoor environments by extracting detailed and task-specific information.

Losses For the loss function, we used Berhu loss, augmented with gradient loss based on Sobel filter detection for all tasks except for semantic segmentation and binary mask prediction. For those tasks, we applied a combination of cross-entropy loss and Dice loss. Given the multi-task nature of our framework, we opted for a linear scalarization objective (equation Equation 1) to efficiently combine the losses from each task. Specifically, we linearly combined all task-specific losses and then applied hypervolume scalarization (Equation 2) to identify the maximum loss. This hybrid objective function (Equation 3) was controlled by a learnable parameter, α, which balanced the influence of the linear scalarization and hypervolume scalarization components:

$$\mathcal{L}_{lin} = \sum_{k=1}^{N} w_k \mathcal{L}_k, \tag{1}$$

$$\mathcal{L}_{hyp} = \max_{k \in [1,N]} \{w_k \mathcal{L}_k\}, \tag{2}$$

$$\mathcal{L}_{hybrid} = \alpha \mathcal{L}_{hyp} + (1-\alpha)\mathcal{L}_{lin}, \tag{3}$$

This hybrid loss formulation allowed for better optimization by dynamically adjusting to the most critical task at each step of the training process.

5. Lighting estimation

For the sake of interactivity and fast computations, we developed a method for estimating lighting parameters using spherical Gaussian (SG) models and we applied them to indoor scenes captured in omnidirectional images. The method involves optimizing light parameters to minimize the residuals between the ground truth and approximated images, thus enabling approximate shading of the scene, and the virtual placement of 3D objects inside the scene with consistent lighting.

Problem Formulation The goal is to estimate the lighting conditions in a given indoor scene represented by an RGB image I, a normal map N, a depth map D, and an albedo map R. The lighting is modeled using a set of spherical Gaussian (SG) functions, each characterized by a set of parameters $\mathbf{x} = x_i$. The optimization problem can be formulated as:

$$\min_{\mathbf{x}} |I - I_a(x, D, N, R))|^2, \tag{4}$$

where I is the ground truth image, and $I_a(x, D, N, R)$ is a function representing the approximate lighting model and depending on the lighting parameters $\mathbf{x}$. We designed the model in a way to be efficient both for the estimation stage, as well as for the real-time rendering component. The estimated lighting parameters $\mathbf{x}$ are then incorporated in a GLSL shader and applied to virtual objects placed inside the scene.

Lighting Model The lighting is modeled using Spherical Gaussian (SG) functions [WRG*09]. This model became recently popular in the graphics community since it can be used for creating differentiable renderers [LWH*22,ZLW*21], and it is very efficient for real time approximations for gaming [HJL*20]. The light intensity of a general SG function is given by:

$$\Gamma(\omega; \sigma, \xi) = e^{\sigma(\omega \cdot \xi - 1)}, \tag{5}$$

where ω is the direction vector, σ is a sharpness parameter, and ξ is the direction of the lobe.

$$\Gamma(\omega; \sigma, \xi) = e^{\sigma(\omega \cdot \xi - 1)}, \tag{6}$$

The Spherical Gaussian formulation is a convenient way to approximate lighting, since it is possible to compute the integral of the SG over the hemisphere as:

$$\int_{\Omega} \Gamma(\omega, \sigma, \xi) d\omega = \frac{2\pi \left(1 - e^{-2\sigma}\right)}{\sigma}, \tag{7}$$

and also other properties useful for shading can be applied. For example, the inner product of two SGs with different parameters $\Gamma_x(\omega; \sigma_x, \xi_x)$ and $\Gamma_y(\omega; \sigma_y, \xi_y)$ is the integral over the hemisphere of the product of the two Gaussians and it is defined in the following way:

$$\int_{\Omega} \Gamma_x(\omega)\Gamma_y(\omega) d\omega = \frac{2\pi e^{d_m - \sigma_x - \sigma_y} \left(1 - e^{-2d_m}\right)}{d_m}, \tag{8}$$

where $d_m = |\sigma_x \xi_x + \sigma_y \xi_y|$.

This formulation enables the approximated computation of specular and diffuse components, given a point in the scene characterized by position $\mathbf{p}$ and normal $\mathbf{n}$. In our formulation, we further consider an additional distance decay parameter to approximate local effects, hence each Gaussian light Γ_L is parametrized by a color C_L, a 3d position μ_L, a 3D direction ξ_L, a sharpness parameter σ_L and a distance decay γ_L.

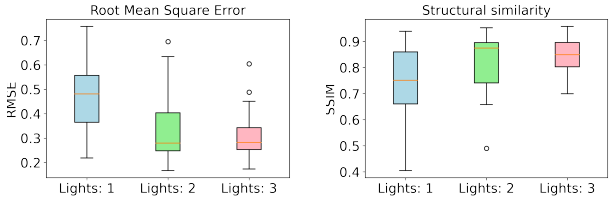

Figure 5: ***Lighting estimation performance.*** *Boxplots of root mean square error (left) and structural similarity(right) with respect to the number of Gaussian light sources.*

Diffuse and specular term In a typical rendering scenario, we have a surface point $\mathbf{p}$ being lit by a light source L, represented by an SG named Γ_L. The outgoing radiance towards the eye for a surface with a Lambertian diffuse BRDF is given by:

$$L(o,\mathbf{p}) = \frac{1}{\pi}\int_{\Omega} L(\omega,\mathbf{p})\cos(\theta_i)\,d\omega, \tag{9}$$

where θ_i is the angle between the incident light direction ω and the surface normal $\mathbf{n}$. If the light is represented by an SG, the latter integral can be approximated as an inner product between an SG and a cosine lobe. In our model, we consider the approximation proposed by Pettineo and Hill and referenced in [Tok22]. In this way, the diffuse irradiance for a given normal $\mathbf{n}$ and the light SG parameters is computed as:

$$\Gamma_{\text{diff}}(\mathbf{n},\sigma,\xi) = \alpha\cdot y+\beta, \tag{10}$$

where α and β are empirically fitted parameters, and y is a clamped value based on the dot product $\xi\cdot\mathbf{n}$. Once approximated, the integral can be computed with Equation 7. On the other side, the specular lighting for a surface point $\mathbf{x}$ due to a light source L represented by an SG Γ_L is computed using the Phong reflection model. The outgoing radiance $L(o,\mathbf{x})$ towards the eye direction $\mathbf{v}$ is given by:

$$L(o,\mathbf{x}) = \int_{\Omega} L(i,\mathbf{x})\,R(i,o,\mathbf{x})\,(\mathbf{n}\cdot\mathbf{i})\,d\Omega, \tag{11}$$

where $L(i,\mathbf{x})$ is the incoming radiance from direction $\mathbf{i}$, $R(i,o,\mathbf{x})$ is the reflectance, and $\mathbf{n}$ is the surface normal at point $\mathbf{x}$. For approximating the specular term we consider again the model proposed by Pettineo and Hill and referenced in [Tok22], based on visbility approximated through the GGX distribution [Hei18, KHDN22] and spherical Gaussian warping.

Lighting estimation We developed an algorithm for estimating Gaussian light parameters, that take as input the inferred signals of the empty scene: the depth, the normal, the semantic, and the material properties in the form of reflectance. The script involves least-squares optimization aiming to minimize the difference between the ground truth image $\mathbf{I}_{\text{gt}}$ and the rendered image $\mathbf{I}_{\text{a}}$ according to the lighting model described above:

$$\mathbf{x}^* = \arg\min_{\mathbf{x}} \sum_{v=1}^{W}\sum_{u=1}^{H} \left|\mathbf{I}_{\text{gt}}[v,u]-\mathbf{I}_{\text{a}}[v,u]\right|^2. \tag{12}$$

The vector $\mathbf{x}$ contains the parameters for each light source, including color, direction, position, spread, and distance decay. We iteratively run the optimization by adding the Gaussian lights one by one, up to a maximum of three lights per scene. To reduce the computation time we perform uniform sampling over the scene (in our experiments we consider 1024 samples), while for the initialization we consider random light positions from a potential set of candidates extracted from a binarized version of the semantic signal, obtained by considering windows, lights, and other potential illuminators. The optimization is performed using the *least_squares* method from the SciPy library, which minimizes the sum of squared residuals with bounds on the parameter values. The output of the algorithm is the set of light parameters that can be used for relighting the scene, and for performing illumination of virtual objects. Fig. 3 shows an example of the effects of adding Gaussian light sources on the iterative approximation of the diffuse map: from left to right the outcomes for Gaussian lights ranging from one to three, with the light maps on the top, and the diffuse maps on the bottom.

Rendering We implemented the Spherical Gaussian lighting model considered for lighting estimation as a GLSL fragment shader inside an OpenGL interactive editing and rendering system [TPG*23] for panoramic images. The framework can import 3D textured assets inside the rasterized version of the empty panoramic scene, and apply affine transformation to place them in different positions of the scene to obtain a restaged scene. The shader can also apply environment mapping.

6. Results

Our framework was implemented in python using *PyTorch* for the processing part, and in OpenGL and Qt for the editing and rendering component. For the experiments, we used a cluster with four Nvidia RTX 4090 for training the multi-task transformer, and a laptop connected to an e-GPU Nvidia RTX Titanium for the lighting estimation and for rendering.

Training data In our experiments, we evaluate the effectiveness of our proposed method using the Structured3D dataset, a synthetic indoor panoramic dataset. Structured3D contains 21,835 panoramic images, each annotated with semantic labels, depth, surface normals, reflectance, and shading information. However, some images had incomplete or corrupted annotations, so we carefully cleaned the dataset, resulting in a final set of 17,434 images. The dataset includes two distinct sets: full-scene and empty-scene images. For our experiments, we use the input RGB images from the full-scene set, while the ground truth for other signals (such as depth, surface normals, and shading) is taken from the empty-scene set. As the original dataset does not provide an official split for training and testing, we adopted the partitioning strategy proposed by [SSP*24].

Training setup We used the Structured3D dataset for both training and testing the model's performance, adhering to the data split outlined in MultiPanoWise [STA*24]. We employed the AdamW optimizer with an initial learning rate of $1e^{-4}$ and set the batch size to one. Both training and evaluation were conducted with an image resolution of 512×1024. The model was trained for up to 30

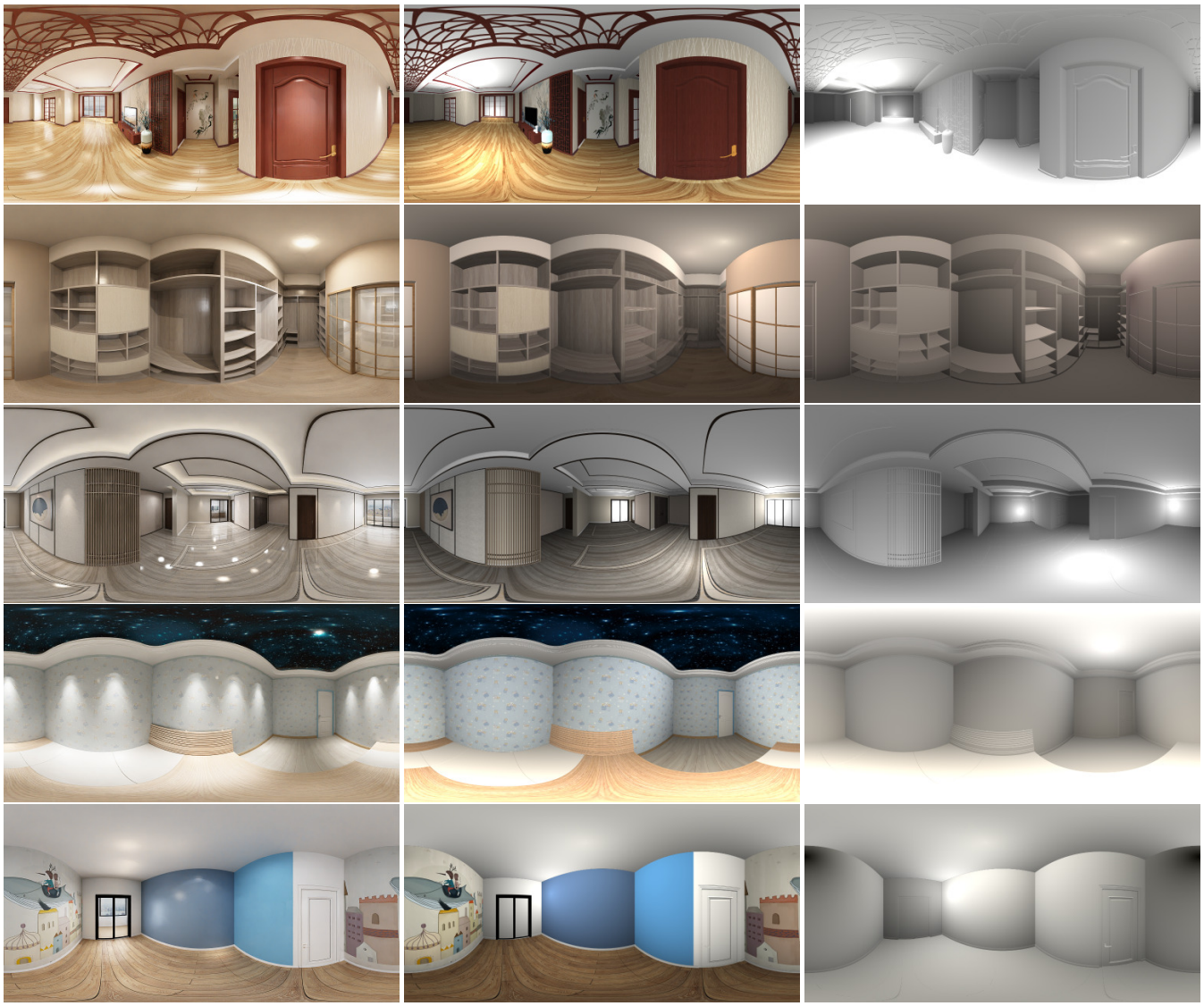

Figure 6: ***Examples of lighting estimation***. *From left to right, the ground truth of the empty scene, the approximated rendering with the Gaussian lighting model, and the corresponding diffuse map. The method is able to capture the general shading appearance of the scene, but it fails to reconstruct local effects due to multiple lights and inter-reflections.*

Approach	Depth			Shading	Normal	Semantic		Albedo		Empty RGB		
	MAE ↓	RMSE ↓	σ_1 ↑	MSE ↓	MAN ↓	mIoU ↑	Dice ↑	MSE ↓	PSNR ↑	MSE ↓	PSNR ↑	SSIM ↑
Baseline [SSP*24]	0.072	0.174	0.952	0.121	7.242	0.701	0.914	0.054	19.15	0.054	20.25	0.826
Pintore et. al. [PAAG22]	0.091	0.197	0.954							0.014	24.70	0.925
Ours	0.057	0.154	0.971	0.116	7.069	0.757	0.939	0.031	22.73	0.023	23.76	0.845

Table 1: ***Quantative comparison of our model with baseline and current SOTA****Presents a quantitative evaluation of various approaches on a specific visual task. The metrics used include Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index (SSIM), Mean Angular error (MAN) and Intersection over Union (IoU). The results are categorized into different aspects of the task, such as depth, shading, normal, semantic, albedo, and empty RGB.*

epochs, with the best-performing weights retained based on validation after each epoch.

Signal estimation results In this section, we present the evaluation results of our multi-task model on the Structured3D dataset, which is used to estimate various signals. To the best of our knowledge, our model is the first to utilize a single panoramic image to generate multiple output signals. Fig. 4 shows the qualitative examples of model prediction vs ground truth.

Figure 7: ***Example of virtual staging****. On the top the original scene versus the restaged one. On the bottom a detail comparison. Note the two armadillo models differently shaded.*

Figure 8: ***Example of virtual staging****. On the top is the original scene versus the restaged one. On the bottom a detail comparison. Note the bunnies naturally integrated in the scene.*

Depth Estimation: We assessed the performance of our depth estimation using metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and σ_1. Table 1 provides a quantitative comparison between our model and current state-of-the-art models. Our model demonstrates significant improvements over recent methods, including the Instant Empty and baseline models. For instance, Instant Empty achieves an MAE of 0.091, whereas our model's MAE is 0.057. Additionally, Instant Empty's RMSE is 0.197 compared to our model's 0.157, and Instant Empty's σ_1 is 0.954, while our model achieves 0.971.

Empty RGB Estimation: We evaluated the performance of our empty RGB estimation using Mean Squared Error (MSE), Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM). Our model demonstrates competitive results relative to the current state-of-the-art. Although the state-of-the-art model

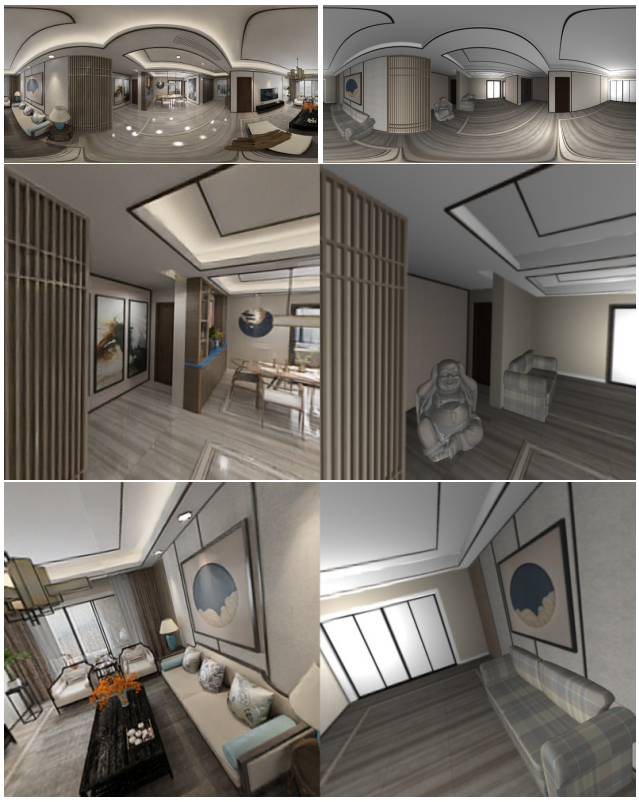

Figure 9: ***Example of virtual staging****. On the top is the original scene versus the restaged one. On the middle and bottom comparisons between details. Note the integration of the couch.*

Figure 10: ***Example of virtual staging****. On the top is the original scene versus the restaged one. On the bottom comparisons between details. Note the different shading of the flowers.*

Figure 11: ***Example of virtual staging****. On the top is the original scene versus the restaged one. On the bottom comparisons between details. Note the dragons differently illuminated.*

achieves a best PSNR of 24.7, our model's best score is 23.76. Despite this, our model outperforms the baseline model.

Semantic Segmentation: We evaluated our semantic segmentation performance using the mean Intersection over Union (mIoU) and Dice score. To the best of our knowledge, no existing model concurrently removes objects from the scene and predicts dense semantic labels for the empty scene. Our model achieves a best mIoU of 0.757 and a Dice score of 0.939, compared to the baseline model's mIoU of 0.701 and Dice score of 0.914.

Shading Estimation: We assessed the performance of shading estimation using Mean Squared Error (MSE). Our model achieves an MSE of 0.116, which is better than the baseline model's MSE of 0.121.

Albedo Estimation: We evaluated the performance of albedo estimation using MSE and PSNR metrics. Our model achieved a PSNR of 22.73 and an MSE of 0.031, whereas the baseline model reached a PSNR of 19.5 and an MSE of 0.054.

Surface Normal Estimation: We assessed performance using the Mean Angular Error (MAN). Our model achieves a mean angular error of 7.069, compared to the baseline model's mean angular error of 7.24.

# Lights	Time (sec)	RMSE	SSIM
1	57.17 $\pm$ 13.52	0.464 $\pm$ 0.149	0.727 $\pm$ 0.157
2	171.1 $\pm$ 40.9	0.335 $\pm$ 0.139	0.818 $\pm$ 0.113
3	429.9 $\pm$ 121.0	0.31 $\pm$ 0.104	0.843 $\pm$ 0.07

Table 2: ***Lighting estimation statistics:*** *average and standard deviation values for processing time (seconds), and for RMSE and SSIM accuracy metrics.*

Lighting estimation results We tested our lighting estimation model on a subset of 20 scenes extracted from the synthetic dataset Structured3D [ZZL*20]. We compared the approximation of lighting with a parametrization involving a different number of lights ranging from one to three. Tab. 2 reports the statistics for lighting estimation: namely the processing time in seconds, and the root mean square error and structural similarity with respect to the ground truth. Fig. 5 shows the boxplots related to the root mean square error and structural similarity with respect to the number of light sources, showcasing a slight improvement for three light sources: we also tried few experiments with a number of lights higher than three without obtaining significant improvement. For a qualitative comparison, we show some examples of the outcomes of lighting estimation in Fig. 6: from left to right, the ground truth image, the approximated one with three Gaussian light sources, and the corresponding diffuse map. From this figure, it is evident that the method is able to reconstruct a plausible estimation of the overall scene, but it fails to reconstruct local effects due to multiple lights and inter-reflections.

Virtual staging We tested the lighting estimation for restaging some scenes extracted from the Structured3D dataset [ZZL*20]. To this end, we used public domain assets for digital furnishing (`https://www.sweethome3d.com/`) and standard models used in CG for testing a rendering algorithm (bunny, dragon, buddha). Fig. 7, Fig. 8, Fig. 9, Fig. 10, and Fig. 11 show examples of restaging operations performed with our editing prototype. From our first initial tests we could see that the illumination of the virtual objects appears plausible and consistent: see the bunnies naturally integrated in the shelf in Fig. 8, or the flowers differently illuminated in Fig. 10, or the dragons differently shaded in Fig. 11. Despite the user interface not being optimized for design purposes (lack of collision detection, limited point-and-click operations, lack of orthogonal views), the generation of each restaged scene took a few minutes of effort.

Limitations Despite the promising initial results, the proposed framework has still some limitations resulting in artifacts that need further research:

- The estimation of material properties is currently limited to a single reflectance signal, that we use for both the diffuse and specular components. A more accurate scene characterization should involve the estimation of distinct reflectance signals as well as roughness [LWH*22, ZCB*22]. We plan to extend our processing pipeline to incorporate those signals;
- Our current lighting model estimates the scene with a constant number of light sources that are the same for the whole omnidirectional scene. While this model provides acceptable global approximations, it is not adequate to represent local effects due to indirect lighting and interreflections, that are particularly evident in indoor scenes. We plan to decompose the scene in small portions where to compute local lighting parameterizations;
- Our current rendering system does not consider shadowing. We plan to incorporate shadow mapping techniques and investigate methods for shadow recovery from panoramic images [JYH*24].

7. Conclusions

We presented a framework for the virtual staging of indoor panoramic images, addressing the challenges of lighting estima-

tion and object insertion in immersive environments. By integrating a multi-task vision transformer and spherical Gaussian lighting models, VISPI enables efficient virtual restaging of cluttered indoor scenes using a single image. Our preliminary results demonstrate the system's capability to generate plausible lighting and object placement, though future work will focus on improving material estimation and handling complex lighting scenarios. Moreover, we plan to test the pipeline on real-world imagery and to integrate it with other tools for performing image-based editing of environments [TRP*23]. The framework provides a promising foundation for applications in real estate, interior design, and other domains requiring realistic virtual environment manipulation.

Acknowledgments This publication was supported by NPRP-S 14th Cycle grant 0403-210132 AIN2 from the Qatar National Research Fund (a member of Qatar Foundation). GP and EG also acknowledge the contribution of the Italian National Research Center in High-Performance Computing, Big Data, and Quantum Computing (Next Generation EU PNRR M4C2 Inv 1.4). The findings herein reflect the work and are solely the responsibility, of the authors.

References

[ACC*23] AI H., CAO Z., CAO Y.-P., SHAN Y., WANG L.: Hrdfuse: Monocular 360deg depth estimation by collaboratively learning holistic-with-regional depth distributions. In *Proc. CVPR* (June 2023), pp. 13273–13282. 2

[CA24] CALI C., AGUS M.: Neuroverse: Immersive exploration of 3d ultrastructural brain reconstructions for education and collaborative analysis. In *Proceedings of the 29th International ACM Conference on 3D Web Technology* (2024), pp. 1–10. 2

[DFB*24] DONG Y., FANG C., BO L., DONG Z., TAN P.: Panocontextformer: Panoramic total scene understanding with a transformer. In *Proc. CVPR* (June 2024), pp. 28087–28097. 2

[dJ23] DA SILVEIRA T. L., JUNG C. R.: Omnidirectional visual computing: Foundations, challenges, and applications. *Computers & Graphics 113* (2023), 89–101. 2

[dSPMLJ22] DA SILVEIRA T. L., PINTO P. G., MURRUGARRA-LLERENA J., JUNG C. R.: 3D scene geometry estimation from 360° imagery: A survey. *ACM Computing Surveys 55*, 4 (2022), 1–39. 2

[EGH21] EINABADI F., GUILLEMAUT J.-Y., HILTON A.: Deep neural models for illumination estimation and relighting: A survey. *Computer Graphics Forum 40*, 6 (2021), 315–331. 2

[GDHG*24] GIROUX J., DASTJERDI M. R. K., HOLD-GEOFFROY Y., VAZQUEZ-CORRAL J., LALONDE J.-F.: Towards a perceptual evaluation framework for lighting estimation. In *Proc. CVPR* (June 2024), pp. 4410–4419. 2

[GHGS*19] GARDNER M.-A., HOLD-GEOFFROY Y., SUNKAVALLI K., GAGNE C., LALONDE J.-F.: Deep parametric indoor lighting estimation. In *Proc. ICCV* (October 2019). 3

[GMS*24] GSAXNER C., MORI S., SCHMALSTIEG D., EGGER J., PAAR G., BAILER W., KALKOFEN D.: Deepdr: Deep structure-aware rgb-d inpainting for diminished reality. In *2024 International Conference on 3D Vision (3DV)* (2024), pp. 750–760. 2

[GSH*19] GARON M., SUNKAVALLI K., HADAP S., CARR N., LALONDE J.-F.: Fast spatially-varying indoor lighting estimation. In *Proc. CVPR* (2019), pp. 6908–6917. 3

[Hei18] HEITZ E.: Sampling the ggx distribution of visible normals. *Journal of Computer Graphics Techniques (JCGT) 7*, 4 (2018), 1–13. 6

[HJL*20] HUO Y., JIN S., LIU T., HUA W., WANG R., BAO H.: Spherical gaussian-based lightcuts for glossy interreflections. *Computer Graphics Forum 39*, 6 (2020), 192–203. 5

[JSN24] JI G., SAWYER A. O., NARASIMHAN S. G.: Virtual home staging and relighting from a single panorama under natural illumination. *Machine Vision and Applications 35*, 4 (2024), 98. 1, 3

[JSZ*21] JIANG H., SHENG Z., ZHU S., DONG Z., HUANG R.: Unifuse: Unidirectional fusion for 360° panorama depth estimation. *IEEE Robotics and Automation Letters 6*, 2 (2021), 1519–1526. 2

[JYH*24] JI X., YANG H., HA C., HOU F., BAI Y.: Shadow recovery based on panorama complex environment. In *Fourth International Conference on Computer Vision and Data Mining (ICCVDM 2023)* (2024), vol. 13063, SPIE, pp. 268–273. 9

[KHDN22] KT A., HEITZ E., DUPUY J., NARAYANAN P. J.: Bringing linearly transformed cosines to anisotropic ggx. 6

[LGY*22] LI Y., GUO Y., YAN Z., HUANG X., YE D., REN L.: Omnifusion: 360 monocular depth estimation via geometry-aware fusion. In *2022 Conference on Computer Vision and Pattern Recognition (CVPR)* (New Orleans, USA, June 2022). 2

[LLM21] LI J., LI H., MATSUSHITA Y.: Lighting, reflectance and geometry estimation from 360° panoramic stereo. In *Proc. CVPR* (2021). 3

[LMF*19] LEGENDRE C., MA W.-C., FYFFE G., FLYNN J., CHARBONNEL L., BUSCH J., DEBEVEC P.: Deeplight: learning illumination for unconstrained mobile mixed reality. In *ACM SIGGRAPH 2019 Talks* (New York, NY, USA, 2019), SIGGRAPH '19, ACM. 3

[LSR*20] LI Z., SHAFIEI M., RAMAMOORTHI R., SUNKAVALLI K., CHANDRAKER M.: Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and svbrdf from a single image. In *Proc. CVPR* (June 2020). 3

[LWH*22] LI Z., WANG L., HUANG X., PAN C., YANG J.: Phyir: Physics-based inverse rendering for panoramic indoor images. In *Proc. CVPR* (June 2022), pp. 12713–12723. 3, 5, 9

[LYO*23] LI Z., YU L., OKUNEV M., CHANDRAKER M., DONG Z.: Spatiotemporally consistent hdr indoor lighting estimation. *ACM Transactions on Graphics 42*, 3 (2023), 1–15. 3

[LZS*23] LIU J., ZHANG Q., SHEN X., WU W., WANG X.: Hybrid prior-based diminished reality for indoor panoramic images. In *Computer Graphics International Conference* (2023), Springer, pp. 388–399. 2

[PAA*21] PINTORE G., AGUS M., ALMANSA E., SCHNEIDER J., GOBBETTI E.: Slicenet: Deep dense depth estimation from a single indoor panorama using a slice-based representation. In *Proc. CVPR* (June 2021), pp. 11536–11545.

[PAAG21] PINTORE G., ALMANSA E., AGUS M., GOBBETTI E.: Deep3DLayout: 3D reconstruction of an indoor layout from a spherical panoramic image. *ACM Trans. Graph. 40*, 6 (2021), 250:1–250:12. 1, 2

[PAAG22] PINTORE G., AGUS M., ALMANSA E., GOBBETTI E.: Instant automatic emptying of panoramic indoor scenes. *IEEE Transactions on Visualization and Computer Graphics 28*, 11 (2022), 3629–3639. 2, 4, 7

[PAG20] PINTORE G., AGUS M., GOBBETTI E.: Atlantanet: inferring the 3d indoor layout from a single 360° image beyond the manhattan world assumption. In *Proc. ECCV* (2020), Springer, pp. 432–448. 1

[PBAG23] PINTORE G., BETTIO F., AGUS M., GOBBETTI E.: Deep scene synthesis of Atlanta-world interiors from a single omnidirectional image. *IEEE TVCG 29* (November 2023). 1

[PGGS16] PINTORE G., GANOVELLI F., GOBBETTI E., SCOPIGNO R.: Mobile mapping and visualization of indoor structures to simplify scene understanding and location awareness. In *Computer Vision – ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8-10 and 15-16, 2016, Proceedings, Part II* (October 2016), Springer, pp. 130–145. 1

[PMG*20] PINTORE G., MURA C., GANOVELLI F., FUENTES-PEREZ L., PAJAROLA R., GOBBETTI E.: State-of-the-art in automatic 3d reconstruction of structured indoor environments. *Computer Graphics Forum 39*, 2 (2020), 667–699. 2

[RAYR22] REY-AREA M., YUAN M., RICHARDT C.: 360MonoDepth: High-resolution 360deg monocular depth estimation. In *CVPR* (2022). 2

[SLK*23] SHINDE Y., LEE K., KIPER B., SIMPSON M., HASANZADEH S.: A systematic literature review on 360° panoramic applications in architecture, engineering, and construction (aec) industry. *Journal of Information Technology in Construction 28* (2023). 1

[SLL*22] SHEN Z., LIN C., LIAO K., NIE L., ZHENG Z., ZHAO Y.: Panoformer: Panorama transformer for indoor 360 depth estimation. In *Computer Vision – ECCV 2022* (Cham, 2022), Springer Nature, pp. 195–211. 1, 2, 4

[SSP*24] SHAH U., SCHNEIDER J., PINTORE G., GOBBETTI E., ALZUBAIDI M., HOUSEH M., AGUS M.: EleViT: exploiting element-wise products for designing efficient and lightweight vision transformers. In *Proc. T4V - IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2024). 3, 4, 6, 7

[STA*24] SHAH U., TUKUR M., ALZUBAIDI M., PINTORE G., GOBBETTI E., HOUSEH M., SCHNEIDER J., AGUS M.: MultiPanoWise: holistic deep architecture for multi-task dense prediction from a single panoramic image. In *Proc. OmniCV - IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)* (2024). 2, 6

[Tok22] TOKUYOSHI Y.: Accurate diffuse lighting from spherical gaussian lights. In *ACM SIGGRAPH 2022 Posters* (New York, NY, USA, 2022), SIGGRAPH '22, ACM. 6

[TPG*23] TUKUR M., PINTORE G., GOBBETTI E., SCHNEIDER J., AGUS M.: Spider: A framework for processing, editing and presenting immersive high-resolution spherical indoor scenes. *Graphical Models 128* (2023), 101182. 6

[TRP*23] TUKUR M., REHMAN A. U., PINTORE G., GOBBETTI E., SCHNEIDER J., AGUS M.: Panostyle: Semantic, geometry-aware and shading independent photorealistic style transfer for indoor panoramic scenes. In *Proc. ICCV* (2023), pp. 1553–1564. 10

[TSH*24] TUKUR M., SCHNEIDER J., HOUSEH M., DOKORO A. H., ISMAIL U. I., DAWAKI M., AGUS M.: The metaverse digital environments: A scoping review of the techniques, technologies, and applications. *Journal of King Saud University-Computer and Information Sciences* (2024), 101967. 2

[WL24] WANG H., LI M.: A new era of indoor scene reconstruction: A survey. *IEEE Access 12* (2024), 110160–110192. 2

[WRG*09] WANG J., REN P., GONG M., SNYDER J., GUO B.: All-frequency rendering of dynamic, spatially-varying reflectance. In *ACM SIGGRAPH Asia 2009 papers*. 2009, pp. 1–10. 5

[WYLL22] WANG G., YANG Y., LOY C. C., LIU Z.: Stylelight: Hdr panorama generation for lighting estimation and editing. In *Proc. ECCV* (2022), Springer, pp. 477–492. 3

[WYS*20] WANG F.-E., YEH Y.-H., SUN M., CHIU W.-C., TSAI Y.-H.: Bifuse: Monocular 360 depth estimation via bi-projection fusion. In *Proc. CVPR* (June 2020). 2

[XSD*13] XU K., SUN W.-L., DONG Z., ZHAO D.-Y., WU R.-D., HU S.-M.: Anisotropic spherical gaussians. *ACM Trans. Graph. 32*, 6 (nov 2013). 4

[YHJ*23] YU H., HE L., JIAN B., FENG W., LIU S.: Panelnet: Understanding 360 indoor environment via panel representation. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (Los Alamitos, CA, USA, jun 2023), IEEE Computer Society, pp. 878–887. 2

[YLW*22] YAN Z., LI X., WANG K., ZHANG Z., LI J., YANG J.: Multi-modal masked pre-training for monocular panoramic depth completion. In *Proc. ECCV* (2022), Springer, pp. 378–395. 2

[YSL*23] YUN I., SHIN C., LEE H., LEE H.-J., RHEE C. E.: Egformer: Equirectangular geometry-biased transformer for 360 depth estimation. In *Proc. ICCV* (October 2023), pp. 6101–6112. 2

[ZCB*22] ZHI T., CHEN B., BOYADZHIEV I., KANG S. B., HEBERT M., NARASIMHAN S. G.: Semantically supervised appearance decomposition for virtual staging from a single panorama. *ACM Transactions on Graphics (TOG) 41*, 4 (2022), 1–15. 1, 3, 9

[ZLW*21] ZHANG K., LUAN F., WANG Q., BALA K., SNAVELY N.: Physg: Inverse rendering with spherical gaussians for physics-based material editing and relighting. In *Proc. CVPR* (June 2021), pp. 5453–5462. 5

[ZZL*20] ZHENG J., ZHANG J., LI J., TANG R., GAO S., ZHOU Z.: Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16* (2020), Springer, pp. 519–535. 2, 3, 9