PanoDR: Spherical Panorama Diminished Reality for Indoor Scenes

Vasileios Gkitsas   Vladimiros Sterzentsenko   Nikolaos Zioulis   Georgios Albanis   Dimitrios Zarpalas

Centre for Research and Technology Hellas, Thessaloniki, Greece

{gkitsasv,vladster,nzioulis,galbanis,zarpalas}@iti.gr

Figure 1: Diminishing the highlighted (red mask) object in indoor spherical panorama images. White lines annotate the scene's layout in panorama and perspective views. Left to right: **i)** masked object to remove, **ii)** pure inpainting result of state-of-the-art methods (top row: RFR [18], bottom row: PICNet [46]), **iii)** perspective view of inpainted region by these methods better shows that they do not necessarily respect the scene's structural layout, **iv)** our panorama inpainting that takes a step towards preserving the structural reality, **v)** perspective view of inpainted region by our model, showing superior results both in texture generation and layout preservation. The results in this figure depict cases where RFR and PICNet provide reasonable structural coherence, and aim at showcasing our model's finer-grained accuracy. Figure 4 presents more qualitative examples where the structural in-coherency of RFR and PICNet is more evident.

Abstract

The rising availability of commercial 360° cameras that democratize indoor scanning, has increased the interest for novel applications, such as interior space re-design. Diminished Reality (DR) fulfills the requirement of such applications, to remove existing objects in the scene, essentially translating this to a counterfactual inpainting task. While recent advances in data-driven inpainting have shown significant progress in generating realistic samples, they are not constrained to produce results with reality mapped structures. To preserve the 'reality' in indoor (re-)planning applications, the scene's structure preservation is crucial. To ensure structure-aware counterfactual inpainting, we propose a model that initially predicts the structure of a indoor scene and then uses it to guide the reconstruction of an empty – background only – representation of the same scene. We train and compare against other state-of-the-art methods on a version of the Structured3D dataset [47] modified for DR, showing superior results in both quantitative metrics and qualitative results, but more interestingly, our approach exhibits a much faster convergence rate. Code and models are available at vcl3d.github.io/PanoDR/.

1. Introduction

The advances in omnidirectional imaging are increasing their adoption in various sectors. The introduction of consumer-grade spherical cameras, or mobile applications stitching moving camera videos into spherical panoramas, can enable new user experiences driven by the holistic capture capacity of 360° cameras. One growing field driven by this holistic imaging property is indoor scene understanding as prominently demonstrated in the pioneering work of PanoContext [45].

PanoContext used a spherical panorama to estimate a room's layout, and has been recently succeeded by a large body of data-driven methods [39, 34, 53, 54] to infer generalized indoor scene layouts from a single monocular 360° image. Indoor layouts represent a coarse geometry representation which can nonetheless be used to reconstruct entire buildings from overlapping captures [30, 29, 31]. In addition, the recent availability of real-world [52, 51, 13] or synthetic [15, 47] spherical depth datasets has enabled finer-grained geometry estimation from panoramas.

Apart from the structural understanding of indoor scenes, 360° images also offer a holistic contextual representation of a scene. Be it either for the detection [6] or segmentation [43] of objects, the combination of contex-

tual and structural understanding can drive next-generation applications offering new forms of interaction. One such application is the Augmented Reality (AR) aided planning of interior spaces, in the context of refurnishing, redecorating or retail, which relies on the simultaneous availability of both structural (layout, geometry) and contextual (object related) information. However, an important shortcoming of AR in such a setting is its limitation to only enhance the real world with virtual elements via visual overlay. Yet in the context of interior (re-)design, the removal of objects is also very relevant, which can be achieved through Diminished Reality (DR). The latter can act either in a complementary manner to AR, allowing it to "replace" objects by first removing (DR), and then overlaying (AR), or by purely erasing them from the scene.

DR is an important technology which has been overlooked for 360° content. It is very closely related to image and/or video inpainting which seeks to reconstruct missing image regions and has achieved impressive results using modern data-driven models. Nevertheless, it also takes a step beyond traditional inpainting as it needs to respect the surrounding context in stricter ways, moving towards realistic reconstructions instead of plausible ones. Indeed, image inpainting state-of-the-art currently focuses solely on photorealism [24], and even the richness of the plausible inpaints [46], foregoing constraints about the alignment of the hallucinated content with the actual scene.

In this work, we focus on 360° DR in the context of interior (re-)design applications, relying on image inpainting and recently available datasets to reconstruct foreground occluded areas faithfully. The contextual realism required to move beyond plausible inpaints in this case is reliant on the scene's structure, a very important cue for indoor space replanning and/or refurnishing that needs to be respected. In summary our contributions are the following:

- We present the first, to the authors' knowledge, method for DR-oriented inpainting in 360° images using recently available panorama datasets with paired full and empty scenes (Figure 2).

- To preserve the structural reality when diminishing scenes we bridge image-to-image translation with generative inpainting, conditioning the inpainting results on the underlying scene structure.

- A DR model with fast convergence rate, significantly faster than state-of-the-art inpainting techniques, outperforming them in both photorealism and structural coherency.

2. Related Work

Data-driven Inpainting. Context Encoders [28], one of the first data-driven methods designed for image inpainting, combined an autoencoder with an adversarial loss in order to generate sharper images. Similarly, Iizuka et al. [9] used two discriminators, one global and one local, to enforce photo-consistency, while also increasing the model's receptive field using dilated convolutions. These early methods conditioned their predictions on both valid and masked inputs, leading to visual artifacts such as color discrepancy and blurriness. To overcome such limitations, Liu et al. [20] introduced partial convolutions for image inpainting to prevent the accumulation of zeros in the encoded representations. Yu et al. [41] extended this idea by proposing gated convolutions to learn the mask automatically which, combined with SN-PatchGAN [23, 16], achieved higher quality inpainting results.

Zeng et al. [42] showed that a pyramid-context encoder exploiting the information of different scales, improved the image completion result. The pyramid-context encoder progressively learned region affinity by attention from a high-level semantic feature map, transferring the learned attention to the low-level feature maps. Another work relying on multi-scale feature fusion was the mutual encoder-decoder work of Liu et al. [21]. CNN features from shallow and deep layers were used to represent the textures and structures of the input image, respectively. Splitting these semantically different but complementary representations into two branches and jointly exploiting them to inpaint in multiple scales produces high quality results. Li et.al [18] propose a recurrent (*i.e.* iterative) inpainting method inspired by how humans inpaint from the outer regions towards the inner ones. They progressively reduce the size of the hole by exploiting the correlation between neighboring pixels and strengthen the constraints for estimating deeper pixels. A Knowledge Consistent Attention module is further utilized that recurrently estimates at each step the attention score for the hole by taking into account the score at the previous step.

As mentioned earlier though, the inpainting task focuses on plausibility and not necessarily the restoration of the real content. This was the main focus of the work of Zheng et al. [46] where a probabilistically principled framework was designed to generate multiple plausible results with reasonable content for each masked input. To achieve that, it combined both generative and variation synthesis approaches. Two generators with shared weights were utilized, with the first taking the masked image as input and sampling the encoding vector from a learned probability distribution, which is subsequently decoded to produce the output image. The second one, used only during training, jointly leverages the masked regions and the feature maps of the first decoder to generate the inpainted result.

Taking into consideration the inherent ambiguity of the image inpainting task, generating plausible realistic images based only on reconstruction losses at pixel ($L1$ loss) or at feature level [11, 5] is not feasible. In cases where novel

contents are missing from the input image, the goal is to generate visual plausible textures, coherent with surrounding known regions while in parallel maintaining the global semantic structure of the image. In order to hallucinate such contents, the generative models' contribution is crucial. The established zero-sum game between the generator and discriminator enforces the former to synthesize crisp images that cannot be distinguished from the natural image distribution. Yet this is accomplished in a pure learning framework, with no additional constraints or guarantees of structural alignment.

Boundary Preserving Inpainting. One important component of photorealism is the preservation of boundaries. In [17] the partial convolution was revisited and used jointly with visual structure reconstruction layers which incorporate structural information in the reconstructed feature map. This resulted in a progressive joint reconstruction of the visual structure (edges) and features in a progressive manner. Also, EdgeConnect [25] introduced an edge generator to hallucinate edges in the missing regions which afterward act as structural guidance for the inpainting task. Likewise, StructureFlow [8] utilize a two-stage network which consists of a structure reconstructor based on [38] and a texture generator. The former produces a smooth image with preserved strong edges, while the latter employs appearance flow [48] to deliver realistic texture. Nevertheless, the edge generator in EdgeConnect discards useful information such as image color whilst StructureFlow uses the input image structure only in the first layer, with no guarantees of structural alignment in the deeper layers.

Image-to-image Translation. Similar to DR inpainting, structure and boundary preservation is very important in the context of image-to-image translation [10], a related synthesis task. Isola et al. [10] introduced *pix2pix*, which utilized an image-conditional GAN for multiple applications such as transforming semantic label to photos, sketches to shoes and Google maps to satellite views. One specific application, reconstructing images from semantic labels [3, 4, 12], especially focuses on preserving the boundaries between different classes. As a result, even the earlier attempts [3] reused the boundary highlighting semantic map in multiple stages within their architectures. Recently, the SPADE blocks [26] employed them within a spatially-adaptive normalization layer to propagate the semantic information throughout the network. Specifically, the activations in normalization layers are modulated by the semantic segmentation map through a spatially adaptive learned transformation. Contrary to *pix2pix*, the proposed method does not normalize the input semantic segmentation mask, thus semantic information is better preserved. Nevertheless, SPADE uses just one style code to control the style of the image without inserting style information throughout the network but only at the beginning. To overcome

this limitation, SEAN [49] embeds one style reference for each semantic class and consequently, leverages the style information in the form of spatially-varying normalization parameters.

In our work, we bring the best of both worlds (photorealism and boundary preservation) in a DR-oriented inpainting task. We exploit the photorealistic generation capacity of generative models in combination with the boundary preserving properties of image-to-image translation task, to replace parts of a 360° image with the occluded background. By integrating layout estimation, our goal is to synthesize content that is conditioned on it, therefore preserving the actual scene's structure, taking a step away from plausibility towards realism.

3. Approach

Our goal is to remove (*i.e.* diminish) objects in spherical panoramas of indoor scenes. On the one hand, the regional nature of our problem is similar to image inpainting, as part of the content needs to be photorealistically hallucinated. On the other hand, its context necessitates counterfactual predictions, specifically to hallucinate occluded areas, drawing away from image inpainting, towards image-to-image translation, where only specific traits/parts of the original image need to be preserved, and others need to be adapted to another domain. More specifically, for indoors DR we consider the task of translating a scene filled with objects/foreground to an empty, background only, scene. At the same time, another requirement imposed by the downstream applications related to interior (re-)planning, is the preservation of *reality*, which, apart from the appearance of the scene, also corresponds to the structure of the scene which needs to be respected when planning changes, as it is largely unchangeable. To address this particular problem, we follow a hybrid approach that will be described in this section, to photorealistically hallucinate occluded content from masked regions, while respecting the underlying scene structure. First, in Section 3.1 we describe the data we use, then in Section 3.2 we provide the details of our hybrid structure-preserving counterfactual generative approach for 360° DR, and, finally, in Section 3.3 we present the model's supervision scheme.

3.1. 360° **Diminished Reality Dataset**

While there exists a variety of datasets of indoor scenes like Matterport3D [2] and Stanford2D3D [1], also including spherical panoramas, they are not suitable for DR. The task that we tackle in this work is fundamentally different from all the previously explored methods for the simple reason that removing an object from an image must be supervised with the occluded content.

Therefore, we employ the Structured3D dataset [47], which provides, among others, photo-realistic spherical

Figure 2: (a) original sample with highlighted object mask, (b) empty room, (c) full room with empty supervision, (d) augmented room with objects, (e) augmented room with empty supervision

panoramas of indoor scenes with 3 different configurations (*empty, simple, and full room*), accompanied by layout and semantic label annotations, making it highly suitable for DR object removal. Considering a normalized indoor scene color image $\mathcal{I}_f \in \mathbb{R}^{W \times H \times 3} \mid 0 \leq \mathcal{I}_f(\mathbf{p}) \leq 1$, with $\mathbf{p} \in \Omega$, and Ω being the image domain of width W and height $H = \frac{W}{2}$, which corresponds to either the simple or full configuration of Structured3D, we seek to remove a foreground object by replacing its appearance with the occluded background. The latter corresponds to the matching scene empty image $\mathcal{I}_b$, which contains a minimal, structure only, WCF (wall–ceiling–floor) representation of the same scene. In addition, Structured3D offers the scene layout's junction positions, which after projecting the connected wireframe reconstruction on the panorama, provide a dense layout WCF segmentation $\mathcal{S} \in \mathbb{N}^{W \times H}$ of the $\mathcal{I}_{f/b}$ scene, with $\mathcal{S}(\mathbf{p}) \in \{1, 2, 3\}$.

To learn the removal of objects, we additionally exploit the semantic labeling of each scene. We randomly pick the largest connected component of one of the available foreground classes as the input object mask $\mathcal{M}$ denoting the region to be diminished. The mask $\mathcal{M}$ is a binary mask with ones in the diminished region and zeros elsewhere, while $\bar{\mathcal{M}}$ is its binary inverse mask. Since the masks are pixel perfect, we calculate their convex to simulate a generic, convex polygon region selection. This way, we can supervise the DR task for the diminishing of an input image $\mathcal{I}_f$ at the region denoted by $\mathcal{M}$, using the background image $\mathcal{I}_b$. However, this straightforward way fails in practice because of conflicting supervision signals. Structured3D is photorealistically rendered using physically based ray-tracing. As a result, the inclusion of foreground introduces light ray bounces, and more importantly, the lights themselves, which are added into the scene as new foreground objects participate into the new image formation process, creating

a photo-inconsistency between $\mathcal{I}_f$ and $\mathcal{I}_b$. To overcome this issue, we perform reverse compositing of selected foreground object classes (*i.e.* excluding lights) into the background. While this is not a perfect solution as incident shadows are lost, it allows for proper training that is not hindered by irregular supervision across the training samples. An illustrative example of this process and the defects it solves is presented in Figure 2.

3.2. Structure-Disentangled DR Model

The DR task blurs the line between image inpainting/completion and image-to-image translation. While image completion is among the top trends in the computer vision research, its standard approach is well formulated. More precisely, either pre-defined shape masks (*e.g.* boxes) are used, or free form ones, to corrupt the image, and then supervise learning with the original image. Apart from the minor mask shape difference, DR predictions are counterfactual as they seek to reconstruct occluded areas, but similar to inpainting, it needs to exploit the context of the scene to diminish it.

Image-to-image translation on the other hand adapts the entire context of an image by translating it to another domain, like photo-to-sketch, or labels-to-image [22, 36]. Usually, only part of the context needs to be preserved, most usually the dominant structure. For our DR case, the masked region needs to be infilled with the occluded context. While this can be considered as another domain, it is nonetheless closer to the original domain that traditional image-to-image translation tasks, which means that it can be partly inferred from its surrounding context, compared to a translated one.

Approach. We employ a hybrid approach taking the masked image $\mathcal{I}_m = \mathcal{I}_f \odot \bar{\mathcal{M}} + \mathbf{1} \odot \mathcal{M}$ as input, with $\odot$ denoting element-wise multiplication and $\mathbf{1} \in \Omega$ is an
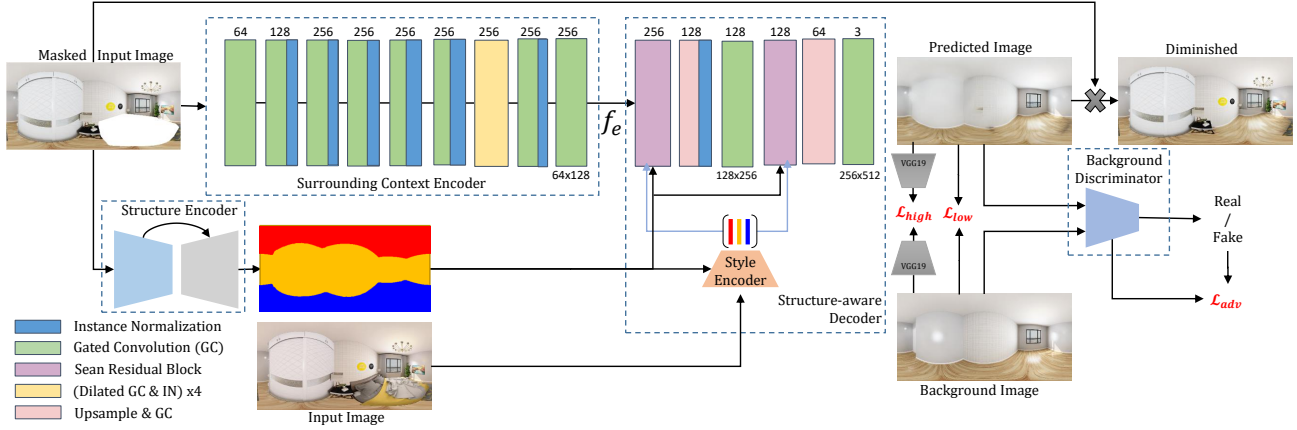
Figure 3: The PanoDR model that preserves the structural reality of the scene while counterfactually inpainting it. The input masked image is encoded twice, once densely by the structure UNet encoder outputing a layout segmentation map, and once by the surrounding context encoder, capturing the scene's context while taking the mask into account via series gated convolutions. These are then combined by the structure-aware decoder with a set of per layout component style codes that are extracted by the complete input image. Two SEAN residual blocks ensure the structural alignment of the reconstructed background image that is supervised by low- and high-level losses, as well as an adversarial loss driven by the background image discriminator. The final diminished result is created via compositing the predicted and input images using the diminishing mask.

all-ones matrix. Our model generates an output image $\hat{\mathcal{I}}_e$ that with is the empty representation of the original input room. The final diminished image $\hat{\mathcal{I}}_d$ is then composited as $\hat{\mathcal{I}}_d = \mathcal{I}_f \odot \bar{\mathcal{M}} + \hat{\mathcal{I}}_e \odot \mathcal{M}$.

To preserve the structural reality, our approach disentangles the structure in an explicit manner, as well as the style, which gets further disentangled per structural element. These are then re-entangled when generating the final background image $\hat{\mathcal{I}}_e$, in order to preserve the structure as faithfully as possible, and to generate the appearance of each structural element distinctly. For image inpainting this is usually done implicitly within the network, even in more recent works [25] that seek to respect the content's structure. For image-to-image translation, boundary information is more important, but the assumption is that the entire image is translated, and that there are no invalid regions like holes that need to be completed.

Structure Encoding. We explicitly encode the scene's structure by predicting a dense layout segmentation map $\hat{\mathcal{S}} \in \Omega$ splitting the panorama into 3 structured regions. For this we use a well established *UNet* segmentation architecture [32], essentially converting structure encoding into a dense classification task. The input to the network is $\mathcal{I}_m$, and it consists of 4 down-sampling and up-sampling convolutional modules joined by skip connections, and uses batch normalization and ReLU activations. The choice of the UNet, with skip connections and 1×1 prediction layers offers finer segmentation results which are very important as they allow for pixel level structure decoupling.

Surrounding Context Encoding. To encode the sur-

rounding context we use an image inpainting derived architecture, which, specifically, is adapted from [41]. Its detailed architecture is depicted in Figure 3, and relies on Gated Convolutions, which are a generalization of partial convolutions [20] that integrates a learnable gating technique when selecting features. In addition, instance normalization [35] and ReLU activations are used. Taking into account the spherical nature of our inputs, we circularly pad [34] in the horizontal image direction all convolution inputs to overcome the longitudinal boundary discontinuity, and use reflection padding to simulate the singularities at the poles [50]. For the bottleneck we rely on repeated dilations [40] to capture the global context more efficiently by expanding the receptive field, avoiding additional parameters and preventing immoderate information loss. This inpainting-derived encoder extracts the content excluding the hole and outputs features $\mathbf{f}_e$.

Structure-aware Decoding. The decoder uses a cascade of SEAN residual blocks [49], gated convolutions and upsample layers. The inputs to the decoder are the predicted structure map $\hat{\mathcal{S}}$, the global context features $\mathbf{f}_e$ as encoded by the surrounding context encoder, and the original – unmasked – image $\mathcal{I}_f$. The latter offers a set of style codes $\mathbf{f}_s^i$ for each structural element $i \in \{wall, ceiling, floor\}$ as learned by a shallow "bottleneck" convolutional autoencoder with a label-wise average pooling output layer. Compared to the global context $\mathbf{f}_e$, these style codes encode local features, corresponding to texture-like details. Through the use of SEAN blocks, we re-entangle the dense global structure $\hat{\mathcal{S}}$, the global surrounding context $\mathbf{f}_e$, and the local style

of each structural element $\mathbf{f}_s^i$. This way, we condition our decoder to respect the global structure while completing the translated image, and modulate the inpainted region's style with the style codes extracted for each structural element.

Background Discriminator. We use a discriminator to adaptively learn the differences between the translated empty images and the corresponding ground truth ones. More specifically, we use a global PatchGAN [10] discriminator with spectral normalization [23]. Its input is either the decoded output $\hat{\mathcal{I}}_e$ or the empty background image $\mathcal{I}_e$ concatenated with the mask $\mathcal{M}$ and classifies each patch of the input image as real or fake. Its output $\mathbf{d}$ is a score map rather than a single score, where each value corresponds to a local region of the input sample covered by its receptive field.

3.3. Supervision

The structure encoding model is supervised with binary cross-entropy, as it is formulated as a dense classification task. For the final inferred background image we use a combination of different domain losses to ensure the photorealistic quality of the predictions:

$$\mathcal{L} = \mathcal{L}_{low} + \mathcal{L}_{high} + \mathcal{L}_{adv}. \tag{1}$$

A low level reconstruction loss $\mathcal{L}_{low}$, a high level synthesis loss $\mathcal{L}_{high}$, and an adaptive adversarial loss $\mathcal{L}_{adv}$.

Low-level Reconstruction Loss. This pixel-based loss focuses on the reconstruction of low frequency components of the predicted image $\hat{\mathcal{I}}_e$:

$$\mathcal{L}_{low} = \lambda_{L1} \frac{1}{N} \sum_{\mathbf{p}}^{\Omega} \mathcal{A} |\mathcal{I}_e - \hat{\mathcal{I}}_e| + \lambda_{tv} \mathcal{M}(|\nabla_x \hat{\mathcal{I}}_e| + |\nabla_y \hat{\mathcal{I}}_e|),$$
$$\tag{2}$$

where N is the total number of pixels, and $\mathcal{A}(\mathbf{p}) \in \mathbb{R}^{W \times H}$ is the spherical attention mask used in [51] that accounts for equirectangular distortion. Apart from the spherically weighted L1 loss, a total variation smoothness prior is used for the diminished area specifically to counter the high frequency artifacts usually seen in the early training stages of generative models.

High-level Synthesis Loss. Apart from encouraging $\hat{\mathcal{I}}_e$ and $\mathcal{I}_e$ to have the same representation at the pixel level with $\mathcal{L}_{low}$, we additionally employ a data-driven loss $\mathcal{L}_{high}$. This enforces them to have a similar representation in the feature space as computed by a CNN model Φ, which in our case, is a pre-trained VGG-19 [33]. Let $\Phi_j(\mathcal{I})$ be the activations of the jth layer of the network Φ, for the given image $\mathcal{I}$, Ω_j its feature element domain, and N_j the total number of feature elements of the j feature map. Then the loss is formulated as a combination of the perceptual and style losses:

$$\mathcal{L}_{high} = \lambda_{perc} \mathcal{L}_{perc} + \lambda_{style} \mathcal{L}_{style} \tag{3}$$

$$\mathcal{L}_{perc} = \sum_j^{P_j} \frac{1}{N_j} \sum_{\boldsymbol{\rho}}^{\Omega_j} |\Phi_j(\mathcal{I}_e) - \Phi_j(\hat{\mathcal{I}}_e)| \tag{4}$$

$$\mathcal{L}_{style} = \sum_j^{S_j} \frac{1}{N_j} \sum_{\boldsymbol{\rho}}^{\Omega_j} \frac{1}{N_j} |\mathcal{G}(\Phi_j(\mathcal{I}_e)) - \mathcal{G}(\Phi_j(\hat{\mathcal{I}}_e))|, \tag{5}$$

where P_j and S_j are the set of features used for the perceptual and style [5, 11] losses, and $\mathcal{G}(\mathbf{M}) = \mathbf{M}\mathbf{M}^T$ is the Gram matrix function. Both losses are derived in a high dimensional data-driven feature space, with the former (perceptual) operating on a global level, and the latter (style) operating on global and local levels.

Adaptive Adversarial Loss. To adaptively improve the quality of the generated background images $\hat{\mathcal{I}}_e$ we additionally employ a discriminator-based loss that is learned during training. Since we use a PatchGAN discriminator, we formulate our combined adversarial loss as a combination of a hinge loss on the final real/fake predictions [19], and a feature matching loss using the discriminator's intermediate features:

$$\mathcal{L}_{adv} = \lambda_D \mathcal{L}_D + \lambda_{FM} \mathcal{L}_{FM} \tag{6}$$

$$\mathcal{L}_D = \frac{1}{N_d} \sum_{\mathbf{p}}^{\Omega_d} r(1 - \mathbf{d}_e) + \frac{1}{N_d} \sum_{\mathbf{p}}^{\Omega_d} r(1 + \mathbf{d}_{\hat{e}}) \tag{7}$$

$$\mathcal{L}_{FM} = \sum_i^{D_i} \frac{1}{N_d^i} \sum_{\mathbf{p}}^{\Omega_d^i} |\mathbf{d}_e^i - \mathbf{d}_{\hat{e}}^i|, \tag{8}$$

where $\mathbf{d}_e$ and $\mathbf{d}_{\hat{e}}$ are the discriminator outputs for the real and predicted background images, Ω_d is the pixel domain of the discriminator's output, N_d the total count of its spatial elements, and the i denotes intermediate discriminator feature maps. The spatial discriminator hinge loss and the feature matching loss are weighted by their respective weights. Feature matching enforces the generator to minimize the statistical difference between the features of the ground truth images and the generated images, which helps further stabilize the training and improve the quality of the generated content.

4. Results

Implementation Details. We implement our model using Pytorch [27] with all experiments conducted on a Nvidia GeForce RTX 3090 GPU. Our generative models are optimized using Adam [14], with $b_1 = 0.5$ and $b_2 = 0.999$, a learning rate of 0.0002 and a batch size of 6. The segmentation UNet is optimized with a default parameterized Adam using a learning rate of 0.0001 and a batch size of 4.

Figure 4: Qualitative results for diminishing objects from scenes in our test set. From left to right: Input image with the diminished area masked with transparent red, RFR, PICNet and ours.

The input and output panorama resolutions are 512×256. The weights of the UNet are initialized with [7] and for the other sub-models from a zero-centered Normal distribution with $\sigma = 0.02$. We empirically set $\lambda_{L1} = 4.0$, $\lambda_{TV} = 1.0$, $\lambda_{perc} = 0.15$, $\lambda_{style} = 40.0$, $\lambda_D = 0.2$ and $\lambda_{FM} = 20.0$.

Table 1: Quantitative results assessing photorealism (LPIPS, PSNR, SSIM, MAE) and structural preservation (mIoU) on the S3D test set.

Method	LPIPS ↓	PSNR ↑	SSIM ↑	MAE ↓	mIoU ↑
RFR [18]	0.0510	31.0114	0.9528	0.0067	0.8583
PICNet [46]	0.0533	32.3072	0.9557	0.0070	0.8502
Ours	**0.0398**	**33.6611**	**0.9620**	**0.0058**	**0.8768**

Experiments. We compare our approach with the following state-of-the-art inpainting methods, PICNet [46] and RFR [18]. We use their official implementations to train both models on our adapted S3D dataset till convergence using the same empty groundtruth images. For our test set

we use 569 images containing objects from the official S3D split, with the diminished regions masked as described in Section 3.1.

Quantitative Comparison. Table 1 compares the photorealistic performance of our model compared to PICNet and RFR. Standard metrics are used, the Mean Absolute Error (MAE), the Peak Signal-to-Noise Ratio (PSNR), the Structural Similarity Index (SSIM) [37], and the Learned Perceptual Image Patch Similarity (LPIPS) [44]. LPIPS is a metric that has been shown to better assess the perceptual similarity between two images. It measures the distance between the target and generated images using features extracted from a pre-trained VGG-16 model. Given the generative nature of our task which aims at natural diminishing/counterfactual inpainting, thus generating plausible and photo-realistic content, LPIPS is considered as an important metric for our evaluation. Nonetheless, our goal is to preserve reality as well, therefore full reference objective measures are also important indicators. We observe that our approach generates content that is perceptually closer to the

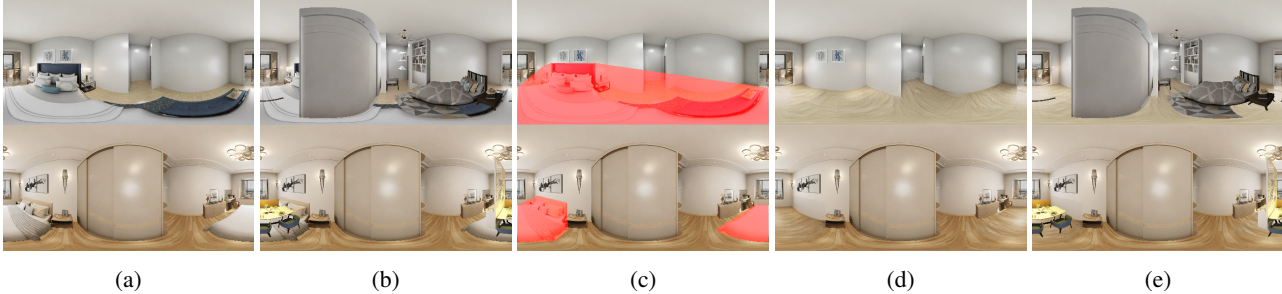|				
(a)	(b)	(c)	(d)	(e)

Figure 5: A demonstration of our method, which levitates AR/DR applications. (a) original panorama (b) augmented reality without diminished reality (c) highlighted object for removal (d) our inpainting method result (e) augmented reality with diminished reality, the result is much more natural.

natural image distribution as shown by the LPIPS metric and the close to 25% performance gain compared to both PICNet and RFR. Furthermore, our method is likely to generate smoother results, which are reflected through the per-pixel and pixel neighborhood based objective metrics.

Moreover, in the DR context, the structural boundary preservation is also of interest. To assess performance with respect to that, we train a layout segmentation model without holes, but with the same training setup as the structure encoder (Section 3.2), and use it to measure the mean intersection-over-union (mIoU) between the groundtruth and the diminished/inpainted results from each respective model. The evaluation is focused on the segmentation results inside the mask $\mathcal{M}$. As indicated in Table 1, our model preserves the structural boundary more consistently than PICNet and RFR alike.

Convergence Analysis. It should also be noted that our approach exhibits significantly faster convergence. The results for PICNet and RFR are trained for 160 and 190 epochs respectively, while our model is only trained for 60 epochs. This is attributed to the interaction between the structure preserving SEAN blocks and the discriminator. Longer trains are directly related to the discriminator's performance and added benefits. The adaptive nature of a learnable adversarial loss helps continuously improve results. This adaptation is phased, first focusing on coarse structure, and progressively adapting to finer details as the generative models learns to consistently output coherent structures. However, with our explicit structure reasoning and integration in the decoder, our adversarial loss – supported by the other low and high level losses – can quickly start discriminating details, allowing the model to converge faster to high quality results.

Qualitative Comparison. Figure 4 presents a set of qualitative results depicting the hallucinated content from our method compared to PICNet and RFR. Our result exhibit photo-realistic textures and structures which are coherent with the background of the images. More specifically, the composite photorealism loss allows our method to generates plausible textures. Further, it is evident in both Fig-

ure 1 (where it is highlighted) and Figure 4, that the explicit SEAN blocks guided by the layout segmentation results better preserve each scene's structural boundaries when counterfactually completing it. In contrast, the compared methods exhibit some flaws, especially at surfaces' boundaries due to the absence of structural guidance. Unsurprisingly, in cases where compared methods generate such artifacts, our method can synthesize realistic textures with plausible structures, driven by the disentangling of the structure, and each structural components style as provided by the structure and style encoders and the SEAN blocks.

DR-enhanced AR. Indoor DR technology is important for supporting AR planning applications. While AR can support the addition of new objects into the scene, it fails to achieve its goal of enhancing the planning experience when seeking to replace objects. For such cases, DR can first be applied on the content that is to be replaced, and then AR can composite the new virtual objects. Figure 5 shows this specific use case where DR technology is highly relevant. For these indoor planning use cases, preserving the scene's structure is very important as it allows for productive interactivity without losing the important context.

5. Conclusion

Diminishing reality is a very challenging task as the hallucination of reality in counterfactual settings is hard to constrain. We show that when considering reality in specific contexts, like indoor planning, sufficient preservation of reality is possible. We combine recent advances in two synthesis tasks and a novel dataset to demonstrate increased performance for spherical panorama based DR. Our approach relies on the segmentation results, which is a limitation given that the fragility of one sub-model is heavily inter-twinned with the resulting diminishing performance. One line of research forward for this task would be complete full-to-empty image translation, which would, nonetheless, require an end-to-end model that would be able to separate the foreground from the background.

References

[1] Iro Armeni, Sasha Sax, Amir R Zamir, and Silvio Savarese. Joint 2d-3d-semantic data for indoor scene understanding. *arXiv preprint arXiv:1702.01105*, 2017. 3

[2] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *7th IEEE International Conference on 3D Vision (3DV)*, pages 667–676. Institute of Electrical and Electronics Engineers Inc., 2018. 3

[3] Qifeng Chen and Vladlen Koltun. Photographic image synthesis with cascaded refinement networks. In *Proceedings of the IEEE international conference on computer vision*, pages 1511–1520, 2017. 3

[4] Hao Dong, Simiao Yu, Chao Wu, and Yike Guo. Semantic image synthesis via adversarial learning. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5706–5714, 2017. 3

[5] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 2, 6

[6] Julia Guerrero-Viu, Clara Fernandez-Labrador, Cédric Demonceaux, and Jose J Guerrero. What's in my room? object recognition on indoor panoramic images. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pages 567–573. IEEE, 2020. 1

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034, 2015. 7

[8] Liu Hongyu, Jiang Bin, Song Yibing, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *Proceedings of the European Conference on Computer Vision*, 2020. 3

[9] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. *ACM Transactions on Graphics (ToG)*, 36(4):1–14, 2017. 2

[10] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 3, 6

[11] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 2, 6

[12] Levent Karacan, Zeynep Akata, Aykut Erdem, and Erkut Erdem. Learning to generate images of outdoor scenes from attributes and semantic layouts. *arXiv preprint arXiv:1612.00215*, 2016. 3

[13] Antonios Karakottas, Nikolaos Zioulis, Stamatis Samaras, Dimitrios Ataloglou, Vasileios Gkitsas, Dimitrios Zarpalas, and Petros Daras. 360° surface regression with a hypersphere loss. In *2019 International Conference on 3D Vision (3DV)*, pages 258–268. IEEE, 2019. 1

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[15] Jin Lei, Xu Yanyu, Zheng Jia, Zhang Junfei, Tang Rui, Xu Shugong, Yu Jingyi, and Gao Shenghua. Geometric structure based and regularized depth estimation from 360° indoor imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[16] Chuan Li and Michael Wand. Precomputed real-time texture synthesis with markovian generative adversarial networks. In *European conference on computer vision*, pages 702–716. Springer, 2016. 2

[17] Jingyuan Li, Fengxiang He, Lefei Zhang, Bo Du, and Dacheng Tao. Progressive reconstruction of visual structure for image inpainting. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5962–5971, 2019. 3

[18] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7760–7768, 2020. 1, 2, 7

[19] Jae Hyun Lim and Jong Chul Ye. Geometric gan. *arXiv preprint arXiv:1705.02894*, 2017. 6

[20] Guilin Liu, Fitsum A Reda, Kevin J Shih, Ting-Chun Wang, Andrew Tao, and Bryan Catanzaro. Image inpainting for irregular holes using partial convolutions. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 85–100, 2018. 2, 5

[21] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. *arXiv preprint arXiv:2007.06929*, 2020. 2

[22] Yongyi Lu, Shangzhe Wu, Yu-Wing Tai, and Chi-Keung Tang. Image generation from sketch constraint using contextual gan. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 205–220, 2018. 4

[23] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018. 2, 6

[24] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2019. 2

[25] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Qureshi, and Mehran Ebrahimi. Edgeconnect: Structure guided image inpainting using edge prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3, 5

[26] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2337–2346, 2019. 3

[27] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017. 6

[28] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei A Efros. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2536–2544, 2016. 2

[29] Giovanni Pintore, Fabio Ganovelli, Ruggero Pintus, Roberto Scopigno, and Enrico Gobbetti. 3d floor plan recovery from overlapping spherical images. *Computational Visual Media*, 4(4):367–383, 2018. 1

[30] Giovanni Pintore, Valeria Garro, Fabio Ganovelli, Enrico Gobbetti, and Marco Agus. Omnidirectional image capture on mobile devices for fast automatic generation of 2.5 d indoor maps. In *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–9. IEEE, 2016. 1

[31] Giovanni Pintore, Ruggero Pintus, Fabio Ganovelli, Roberto Scopigno, and Enrico Gobbetti. Recovering 3d existing-conditions of indoor structures from spherical images. *Computers & Graphics*, 77:16–29, 2018. 1

[32] 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. 5

[33] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 6

[34] Cheng Sun, Chi-Wei Hsiao, Min Sun, and Hwann-Tzong Chen. Horizonnet: Learning room layout with 1d representation and pano stretch data augmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1047–1056, 2019. 1, 5

[35] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Instance normalization: The missing ingredient for fast stylization. *arXiv preprint arXiv:1607.08022*, 2016. 5

[36] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 4

[37] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 7

[38] Li Xu, Qiong Yan, Yang Xia, and Jiaya Jia. Structure extraction from texture via relative total variation. *ACM transactions on graphics (TOG)*, 31(6):1–10, 2012. 3

[39] Shang-Ta Yang, Fu-En Wang, Chi-Han Peng, Peter Wonka, Min Sun, and Hung-Kuo Chu. Dula-net: A dual-projection network for estimating room layouts from a single rgb panorama. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3363–3372, 2019. 1

[40] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017. 5

[41] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4471–4480, 2019. 2, 5

[42] Yanhong Zeng, Jianlong Fu, Hongyang Chao, and Baining Guo. Learning pyramid-context encoder network for high-quality image inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1486–1494, 2019. 2

[43] Chao Zhang, Stephan Liwicki, William Smith, and Roberto Cipolla. Orientation-aware semantic segmentation on icosahedron spheres. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3533–3541, 2019. 1

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 7

[45] Yinda Zhang, Shuran Song, Ping Tan, and Jianxiong Xiao. Panocontext: A whole-room 3d context model for panoramic scene understanding. In *European conference on computer vision*, pages 668–686. Springer, 2014. 1

[46] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1438–1447, 2019. 1, 2, 7

[47] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Proceedings of The European Conference on Computer Vision (ECCV)*, 2020. 1, 3

[48] Tinghui Zhou, Shubham Tulsiani, Weilun Sun, Jitendra Malik, and Alexei A Efros. View synthesis by appearance flow. In *European conference on computer vision*, pages 286–301. Springer, 2016. 3

[49] Peihao Zhu, Rameen Abdal, Yipeng Qin, and Peter Wonka. Sean: Image synthesis with semantic region-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5104–5113, 2020. 3, 5

[50] Nikolaos Zioulis, Federico Alvarez, Dimitrios Zarpalas, and Petros Daras. Single-shot cuboids: Geodesics-based end-to-end manhattan aligned layout estimation from spherical panoramas. *arXiv preprint arXiv:2102.03939*, 2021. 5

[51] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, Federico Alvarez, and Petros Daras. Spherical view synthesis for self-supervised 360° depth estimation. In *2019 International Conference on 3D Vision (3DV)*, pages 690–699. IEEE, 2019. 1, 6

[52] Nikolaos Zioulis, Antonis Karakottas, Dimitrios Zarpalas, and Petros Daras. Omnidepth: Dense depth estimation for indoors spherical panoramas. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 448–465, 2018. 1

[53] Chuhang Zou, Alex Colburn, Qi Shan, and Derek Hoiem. Layoutnet: Reconstructing the 3d room layout from a single rgb image. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2051–2059, 2018. 1

[54] Chuhang Zou, Jheng-Wei Su, Chi-Han Peng, Alex Colburn, Qi Shan, Peter Wonka, Hung-Kuo Chu, and Derek Hoiem. 3d manhattan room layout reconstruction from a single 360 image. *arXiv preprint arXiv:1910.04099*, 2019. 1