VIEWPORT-ORIENTED PANORAMIC IMAGE INPAINTING

Zhuoyi Shang[⋆†], *Yanwei Liu*[⋆*], *Guoyi Li*[⋆†], *Yunjian Zhang*[⋆†], *Jingbo Miao*[⋆†], *Jinxia Liu*[‡], *Liming Wang*[⋆]

[⋆] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[†] School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
[‡] Zhejiang Wanli University, Ningbo, China

ABSTRACT

Panoramic images are usually viewed through Head Mounted Displays (HMDs), which renders only a narrow field of view from the raw panoramic image. This distinctive viewing feature has largely been ignored when inpainting panoramic images. To address this issue, we propose a viewport-oriented generative adversarial panoramic image inpainting network in this paper. For capturing the distorted features accurately in the generating process of equirectangular projection (ERP) panoramic image, a latitude-adaptive feature fusion module is devised to aggregate the latitude-level features in ERP image and less-distorted patch-level viewport-domain features. Furthermore, a novel cross-domain discriminator is proposed to force the inpainting network to generate more plausible results in viewports. Extensive experiments show that our model achieves better performance compared to the baseline methods, especially in the viewport images.

***Index Terms*—** panoramic image, virtual reality, image inpainting

1. INTRODUCTION

Panoramic images (PIs) have attracted much attention in recent years since they can represent the omnidirectional visual contents. With an increasing need of processing and editing PIs in our daily life, inpainting has become an important task in various PI applications, such as privacy protection in panoramas [1], old panoramic photo restoration, and also PI post-processing for AR [2].

Early image inpainting methods devote to copying pixels from the unmasked regions to fill in holes, but they are inefficient for images with large holes [3]. Applying deep convolutional neural networks (CNNs) [4] to image inpainting problems has made great progress in recent years. Typically, aiming for irregular holes inpainting, LBAM [5] and Contextual Residual Aggregation(CRA) [6] improve the Generative Adversarial Network (GAN) with attention module. However, these methods are originally designed targeting the planar perspective images and perform poorly on PIs. The contents in PIs are usually deformed during the equirectangular projection (ERP) from the raw spherical signal [7] and correspondingly the conventional CNNs fail to extract features from them accurately.

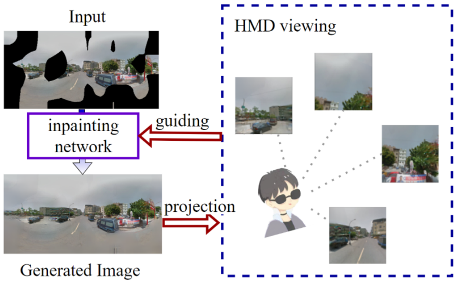


Fig. 1. Our inpainting network aims to generate images with more realistic results in viewports besides the ERP image.

Recently, several pioneer studies were proposed for panoramic image/video inpainting. Some of them improve the off-the-shelf methods for PIs by applying either semantic conditioning [2, 8] or depth information [1]. Besides, some researchers are interested in exploring other projection formats (such as Cubemap in PIINET [9]) to introduce less distortion. PIs are usually viewed with freely changeable viewports via Head Mounted Displays (HMDs). The above-mentioned methods do not take this viewing feature into account, and the generated image may also work poorly in HMDs.

Considering the viewing process of PIs, we propose a viewport-oriented generative adversarial PI inpainting network. An abstract representation of viewport-oriented inpainting is illustrated in Fig. 1. Specifically, we first introduce a latitude-adaptive feature fusion module to guide the generator to capture and generate discriminative features adaptively. The module is designed with a dual-layer structure to fuse the latitude-level features in ERP image and less-distorted patch-level viewport-domain features. Then, we further suggest a cross-domain discriminator to urge the network to generate more desirable results towards viewports besides the ERP

*Corresponding author. This work was supported in part by National Natural Science Foundation of China under Grant 61771469, and the Cooperation project between Chongqing Municipal undergraduate universities and institutes affiliated to CAS (HZ2021015).

image. Extensive experiments are performed with varying ratio of the masked area, and the results show that our method is superior to the existing methods.

2. METHOD

For the previous inpainting methods, the imperfect inpainting noises in ERP image will be transferred to the viewports during viewing via HMDs. Due to the affection of viewport rendering, the inpainting noise in ERP images may be magnified in the viewports. To achieve more satisfying results in both ERP and viewport images, we introduce the viewport information to both the generator and discriminator to optimize the inpainting GAN.

2.1. Overview

The proposed PI inpainting pipeline is shown in Fig. 2. It repairs the damaged image I_{input} with GAN including a generator G to predict the inpainted image I_{out} and a cross-domain discriminator to distinguish the authenticity of repaired image I_{re}. The repaired image I_{re} is the result of replacing the hole region (masked as M) using the generated image I_{out},

$$I_{re} = I_{out} \otimes M + I_{input} \otimes (1 - M), \tag{1}$$

where $\otimes$ denotes the element-wise multiplication operation.

In the generator, latitude-adaptive feature fusion module aggregates the features from both the ERP patches $\{B_{ij}\}$ and their corresponding viewport patches $\{A_{ij}\}$ to generate the implicit information map IIM. IIM connects the symmetrical layers in encoder and decoder of generator, providing available gradients so that the network parameters can be optimized through back-propagation. In the cross-domain discriminator, the authenticity of I_{re} is measured in both ERP domain (D_{ERP} sub-discriminator) and viewport domain (D_V sub-discriminator).

2.2. Latitude-Adaptive Feature Fusion

Aiming for extracting effective features that are rarely affected by the distortion in ERP image, the latitude-adaptive feature fusion module is designed in two tiers: the latitude-level feature extraction (LLFE) is used to describe the latitude-adaptive dependency while the patch-level feature extraction (PLFE) aims at building a bridge connecting ERP features and viewport patches.

The detailed building blocks for the dual-layer design are shown in Fig. 2. To model the long-range latitude-dependent contextualization, LLFE joins viewport features GV_i in the same latitude i and the deep semantic features hf of I_{input}. hf is the encoder layer group of the generator. To aggregate local information more precisely, we establish the correspondence between ERP image patches $\{B_{ij}\}$ and viewport patches $\{A_{ij}\}$ by PLFE.

Patch-level feature extraction. Firstly, the ERP image I_{input} is divided evenly into several patches $\{B_{ij}\}$. For each patch B_{ij}, taking the center point (i, j) as the viewport rendering center, we get the patch group $\{A_{ij}\}$ through viewport projection with 90° field of view(FOV). In Fig. 2, this procedure is called E-to-V projection. We leverage the pre-trained VGG-16 network as the backbone to extract effectively semantic visual information $\{p_{ij}\}$ and $\{q_{ij}\}$,

$$p_{ij} = VGG16(B_{ij}), q_{ij} = VGG16(A_{ij}) \tag{2}$$

For each location in $i \in (0, n-1)$ and $j \in (0, m-1)$, the patch-level implicit information PF_{ij} is calculated between p_{ij} and q_{ij},

$$PF_{ij} = \psi(p_{ij}, q_{ij}) \tag{3}$$

where n and m are the numbers of latitudes and longitudes in the PI, respectively. In the implementation, we chose $m = 4$ and $n = 4$ to cover the whole sphere, and ψ is a calculation function realized by elaborate Gated Convolution(GC).

Latitude-level feature extraction. For latitude $i \in (0, n-1)$, the latitude-level feature value LF_i is calculated by feature groups $GV_i = \{q_{i1}, q_{i2}, ...q_{im}\}$ in latitude i and the feature maps of the whole ERP image hf ,

$$LF_i = \Phi(GV_i, hf) \tag{4}$$

where Φ is a GC function.

Feature fusion. At the pixel position with latitude $i \in (0, n-1)$ and longitude $j \in (0, m-1)$, the results of LLFE LF_i and PLFE PF_{ij} are finally combined as IIM_{ij},

$$IIM_{ij} = \omega(LF_i, PF_{ij}), \tag{5}$$

where ω is a GC function. To make the parameter-updating automatically, we add IIM to the high-level decoding layers in generator for training.

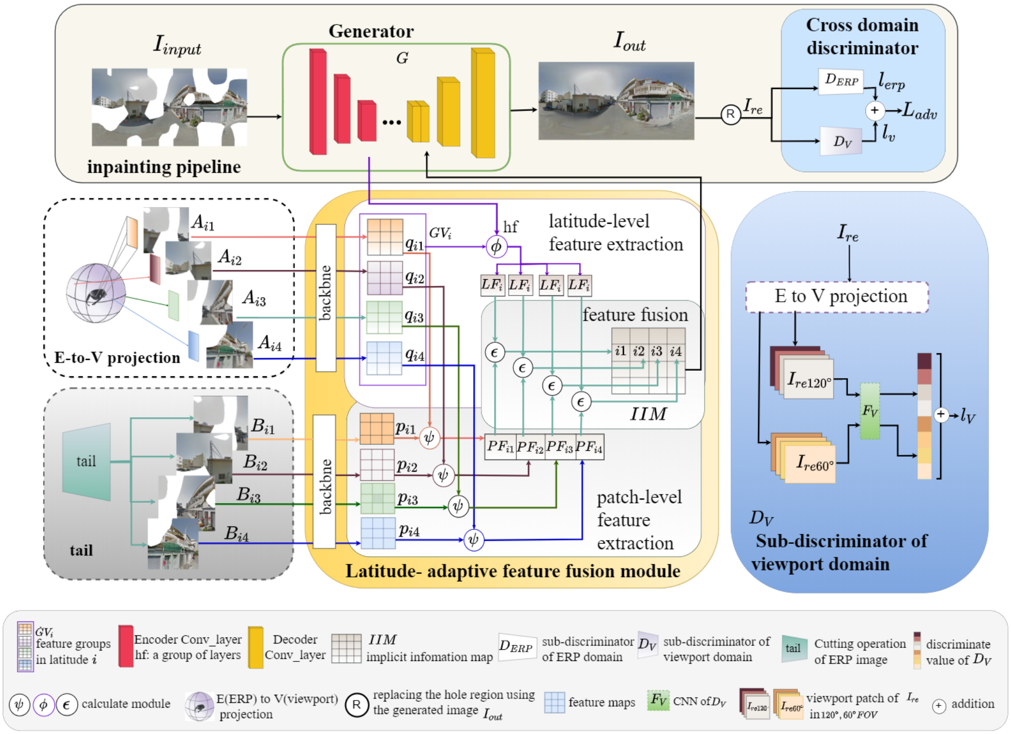
2.3. Cross-domain Discriminator

To obtain more reasonable result, the discriminator is designed in a cross-domain manner: D_{ERP} is utilized to assist in generating images that conform to the distribution characteristics of ERP images and D_V aims to strengthen the model to generate images with better visual quality in HMDs. For the output of the discrimination path, the adversarial generating loss L_{adv} is defined as the combination of ERP adversarial loss l_{ERP} and viewport adversarial loss l_V,

$$L_{adv} = \lambda \cdot l_{ERP} + (1 - \lambda) \cdot l_V \tag{6}$$

where λ is the tradeoff parameter. In our implementation, we empirically set λ = 0.5. Following WGAN-GP [10], we formulate the l_{ERP} as,

$$\begin{aligned} l_{ERP} = \underset{G}{max} \underset{D_{ERP}}{min} & E_{I_{gt} \sim pdata(I_{gt})} D_{ERP}(I_{gt}) \\ & - E_{I_{re} \sim pdata(I_{re})} D_{ERP}(I_{re}) \\ & + \lambda_0 E_{\hat{I} \sim pdata(\hat{I})}((\nabla || D_{ERP}(\hat{I}) ||)^2)^2 \end{aligned} \tag{7}$$



I_{input}
Generator
G
I_{out}
Cross domain discriminator
D_{ERP}
l_{erp}
L_{adv}
D_V
l_v
R
I_{re}
inpainting pipeline
A_{i1}
A_{i2}
A_{i3}
A_{i4}
E-to-V projection
backbne
GV_i
q_{i1}
q_{i2}
q_{i3}
q_{i4}
hf
ϕ
latitude-level feature extraction
LF_i
feature fusion
ϵ
$i1$ $i2$ $i3$ $i4$
IIM
PF_{i1} PF_{i2} PF_{i3} PF_{i4}
p_{i1}
p_{i2}
p_{i3}
p_{i4}
ψ
patch-level feature extraction
tail
B_{i1}
B_{i2}
B_{i3}
B_{i4}
Latitude- adaptive feature fusion module
I_{re}
E to V projection
$I_{re120°}$
$I_{re60°}$
F_V
l_V
D_V
Sub-discriminator of viewport domain
GV_i feature groups in latitude i
Encoder Conv_layer hf: a group of layers
Decoder Conv_layer
IIM implicit infomation map
D_{ERP} sub-discriminator of ERP domain
D_V sub-discriminator of viewport domain
tail Cutting operation of ERP image
discriminate value of D_V
ψ ϕ ϵ calculate module
E(ERP) to V(viewport) projection
R replacing the hole region using the generated image I_{out}
feature maps
F_V CNN of D_V
I_{re120} $I_{re60°}$ viewport patch of I_{re} in 120°, 60° FOV
+ addition

Fig. 2. Viewport-oriented PI inpainting pipeline. The damaged PI in ERP format is inputted into Generator. During training, the latitude-adaptive feature fusion module generates the implicit information map IIM to connect encoder layers and decoder layers in Generator. Repaired image I_{re} is fed to the cross-domain discriminator for adversarial training.

where I_{gt} denotes the ground truth image. $\hat{I}$ is sampled from the distribution of I_{gt} and I_{re} by linear interpolation, and λ_0 is set to 10 in our implementation.

Usually, humans have the highest vision acuity in 60° FOV, and the common HMDs cover 120° FOV [11]. Thus in D_V, we make a coarse and fine-grained multi-FOV information extraction via combining 60° FOV I_{re60° and 120° FOV I_{re120°. The details of D_V are shown in the block in bottom-right corner in Fig. 2, The viewport features are extracted by a co-trained CNN F_V. The viewport adversarial loss l_V is defined as,

$$l_V = \frac{1}{2}(||\sum(F_V(I_{re60^\circ}) + F_V(I_{re120^\circ})) * L_{m\times n} - \sum(F_V(I_{gt60^\circ}) + F_V(I_{gt120^\circ})) * L_{m\times n}||_1) \tag{8}$$

where n and m are the numbers of latitudes and longitudes selected for viewport rendering, respectively, and $L_{m\times n}$ is the set of $m \times n$ learnable parameters.

3. EXPERIMENTS AND DISCUSSION

3.1. Implementation Details

We evaluate our inpainting network on two datasets: 360° StreetView [12] (360-SP) and 3D60 [13, 14, 15]. 3012 images in 3D60 and 2017 images in 360-SP are used as training data. Both the training images and testing images are resized to 512×256. We implement our model using Pytorch and all experiments are conducted on Nvidia TiTanXp GPU. Our model is optimized using Adam optimizer. The learning rates of generator, sub-discriminator D_{ERP}, sub-discriminator D_V are 0.0001, 0.0001 and 0.00001, respectively.

3.2. Comparison with other methods

We compare our model with the following GAN-based baselines: the planar image inpainting models Contextual Residual Aggregation(CRA) [6] and LBAM [5], and the PI model PIINET [9]. All the baseline models are re-trained separately on 3D60 and 360-SP datasets.

Fig. 3 shows the qualitative comparison results, (c) and (d) are the viewports corresponding to the green box of (a) and (b). The result of PIINET is obviously the worst among all methods, which is limited by its unstable training process. Surpassing the PIINET, LBAM still generates blurry results especially in 360-SP datasets, and over-smoothing results are undesirable for viewport viewing. Similarly, CRA produces discontinuous texture in both Figs. 3(a) and (b). In contrast, our model repairs the damaged roof with windows perfectly in Figs. 3(a) and also obtains reasonable results in Fig. 3(b).

Table 1. SSIM and PSNR comparison on 3D60 and 360-SP datasets

Dataset		ERP domain								Viewport domain							
		SSIM				PSNR(dB)				SSIM				PSNR(dB)			
	Mask	0.2	0.3	0.4	0.5	0.2	0.3	0.4	0.5	0.2	0.3	0.4	0.5	0.2	0.3	0.4	0.5
3D60	LBAM [5]	0.92	0.84	0.73	0.53	32.62	28.03	23.69	19.20	0.91	0.84	0.78	0.49	40.28	32.38	28.67	19.59
	CRA[6]	0.93	0.85	0.72	0.59	33.49	26.83	23.35	20.92	0.93	0.84	0.79	0.54	42.15	30.12	28.31	22.51
	PIINET[9]	0.68	0.61	0.51	0.32	20.68	17.54	14.85	11.79	0.67	0.63	0.53	0.35	24.48	21.13	17.13	13.08
	Ours	**0.94**	**0.88**	**0.79**	**0.62**	**36.02**	**30.20**	**26.07**	**22.20**	**0.95**	**0.88**	**0.83**	**0.61**	**43.57**	**35.86**	**32.17**	**24.27**
360-SP	LBAM [5]	**0.93**	0.85	0.74	0.54	33.44	28.35	24.99	21.23	0.92	0.86	0.73	0.51	38.66	31.25	27.65	21.35
	CRA[6]	0.78	0.71	0.61	0.44	25.51	24.28	21.94	19.23	0.68	0.66	0.56	0.38	28.01	27.24	22.70	18.83
	PIINET[9]	0.89	0.81	0.67	0.44	22.76	19.62	16.58	14.03	0.93	0.86	0.69	0.48	23.12	23.88	17.77	14.49
	Ours	**0.93**	**0.86**	**0.77**	**0.57**	**34.33**	**28.75**	**25.46**	**21.43**	**0.95**	**0.88**	**0.74**	**0.52**	**39.19**	**32.09**	**28.36**	**22.37**

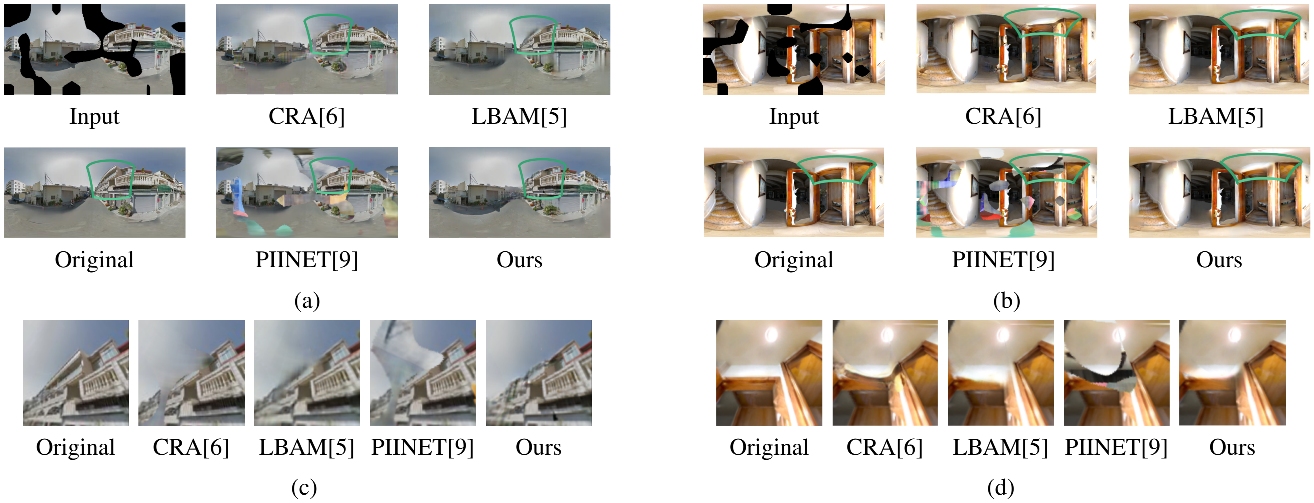


Fig. 3. Visual comparison of our model with baselines. Examples of inpainted images on (a) 360-SP and (b) 3D60 dataset, and the corresponding viewports (c) and (d).

Further, it achieves the better visual quality of viewports in Figs. 3(c) and (d) than the other baselines.

We also compare our model quantitatively with baselines with mask ratios from 0.2 to 0.5, as shown in Table 1. We randomly select three viewports with 90° FOV for viewport quality evaluation. Table 1 shows that CRA performs well in 3D60, but the inpainting performances degrades for 360-SP. Among all the baselines, PIINET performs poorly. Comparably, our model behaves favorably both in two datasets. In specific, the SSIM for viewport domain of 3D60 of our method is higher than LBAM with 12 percent when mask ratio is 0.5.

Table 2. Ablation study on 360-SP datasets

	ERP domain				Viewport domain			
	SSIM		PSNR(dB)		SSIM		PSNR(dB)	
Mask	0.2	0.4	0.2	0.4	0.2	0.4	0.2	0.4
(a)	0.90	0.67	24.86	18.22	0.91	0.70	30.03	18.23
(b)	0.92	0.74	31.91	23.98	0.91	0.74	37.33	25.00
(c)	0.93	0.76	34.40	25.25	0.93	0.75	39.11	25.63

3.3. Ablation Study

We conduct the ablation study on the 360-SP dataset. We modify the second stage generative network(SSGN) in [6] as our generative network. We decompose our algorithm into the sub-algorithms based on the existence of (a) SSGN, (b) SSGN with latitude-adaptive feature fusion module, and (c) SSGN with both latitude-adaptive feature fusion module and cross-domain discriminator. Experimental results for both ERP and viewport images are shown in Table 2. It can be seen that the latitude-adaptive feature fusion module plays an important role in generating more accurate results. The two metric values for viewports are further improved when adding cross-domain discriminator, showing that it provides the effective viewport-oriented guidance.

4. CONCLUSION

In this paper, we propose a viewport-oriented PI inpainting network that enables more reasonable results for viewports. Unlike other PI inpainting methods, we solve this problem with fully consideration of the distortion from PI transferring to viewports. By introducing the latitude-adaptive feature fusion module, our network is effective in capturing and generating the less-distorted features. Furthermore, the cross-domain discriminator is presented to guide the generative network for better results. Qualitative and quantitative experiments show that our method can obtain superior results especially in viewports.

5. REFERENCES

[1] Ries Uittenbogaard, Clint Sebastian, Julien Vijverberg, Bas Boom, Dariu M Gavrila, et al., "Privacy protection in street-view panoramas using depth and multi-view imagery," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10581–10590.

[2] Vasileios Gkitsas, Vladimiros Sterzentsenko, Nikolaos Zioulis, Georgios Albanis, and Dimitrios Zarpalas, "Panodr: Spherical panorama diminished reality for indoor scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 3716–3726.

[3] Avisek Lahiri, Arnav Kumar Jain, Sanskar Agrawal, Pabitra Mitra, and Prabir Kumar Biswas, "Prior guided gan based semantic inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13696–13705.

[4] 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*, 2016, pp. 2536–2544.

[5] Chaohao Xie, Shaohui Liu, Chao Li, Ming-Ming Cheng, Wangmeng Zuo, Xiao Liu, Shilei Wen, and Errui Ding, "Image inpainting with learnable bidirectional attention maps," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 8858–8867.

[6] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu, "Contextual residual aggregation for ultra high-resolution image inpainting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 7508–7517.

[7] Yu-Chuan Su and Kristen Grauman, "Kernel transformer networks for compact spherical convolution," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9442–9451.

[8] Naofumi Akimoto, Seito Kasai, Masaki Hayashi, and Yoshimitsu Aoki, "360-degree image completion by two-stage conditional gans," in *2019 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2019, pp. 4704–4708.

[9] Seo Woo Han and Doug Young Suh, "Piinet: A 360-degree panoramic image inpainting network using a cube map," *arXiv preprint arXiv:2010.16003*, 2020.

[10] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville, "Improved training of wasserstein gans," *Advances in neural information processing systems*, vol. 30, 2017.

[11] Fang-Yi Chao, Lu Zhang, Wassim Hamidouche, and Olivier Déforges, "A multi-fov viewport-based visual saliency model using adaptive weighting losses for 360° images," *IEEE Transactions on Multimedia*, 2020.

[12] Shih-Hsiu Chang, Ching-Ya Chiu, Chia-Sheng Chang, Kuo-Wei Chen, Chih-Yuan Yao, Ruen-Rone Lee, and Hung-Kuo Chu, "Generating 360 outdoor panorama dataset with reliable sun position estimation," in *SIGGRAPH Asia 2018 Posters*, pp. 1–2. 2018.

[13] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang, "Matterport3d: Learning from rgb-d data in indoor environments," *arXiv preprint arXiv:1709.06158*, 2017.

[14] 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.

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