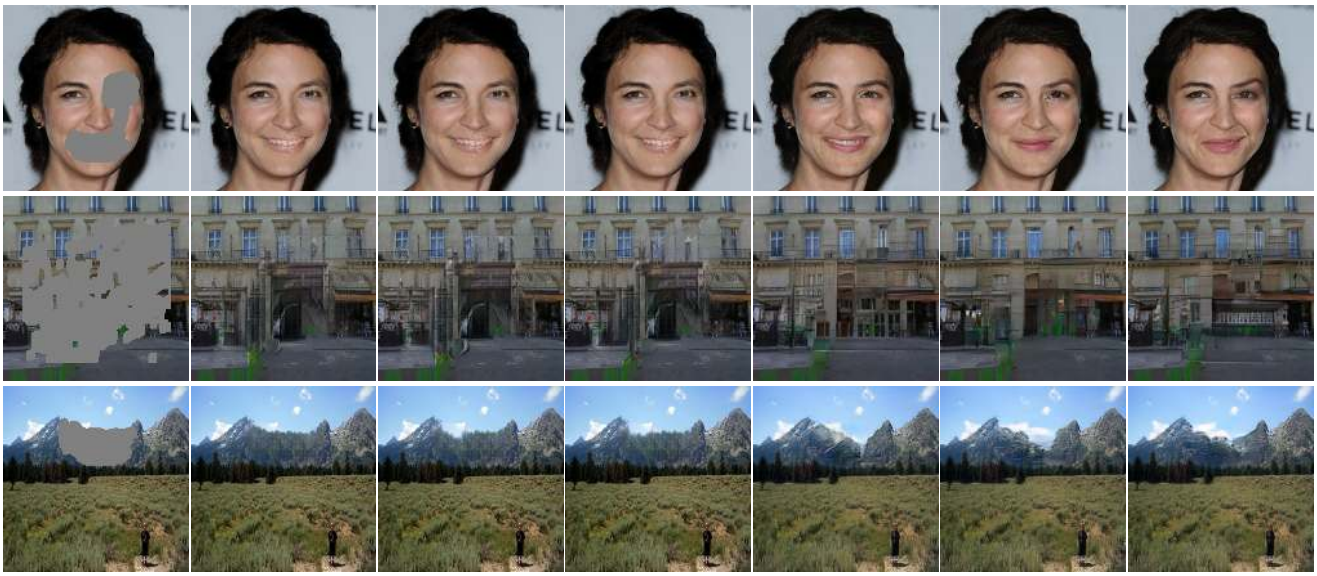
PD-GAN: Probabilistic Diverse GAN for Image Inpainting

Hongyu Liu[1] Ziyu Wan[2] Wei Huang[3] Yibing Song[4] Xintong Han[1*] Jing Liao[2]
[1]Huya Inc [2]City University of Hong Kong [3]Hunan University [4]Tencent AI Lab
{liuhongyu1,hanxintong}@huya.com ziyuwan2-c@my.cityu.edu.hk
yibingsong.cv@gmail.com jingliao@cityu.edu.hk



(a) Input (b) PIC$_1$ [58] (c) PIC$_2$ (d) PIC$_3$ (e) Ours$_1$ (f) Ours$_2$ (g) Ours$_3$

Figure 1. Visual comparison of diverse inpainting results. Our method generates more diverse and visually realistic results than PIC [58].

Abstract

*We propose PD-GAN, a probabilistic diverse GAN for image inpainting. Given an input image with arbitrary hole regions, PD-GAN produces multiple inpainting results with diverse and visually realistic content. Our PD-GAN is built upon a vanilla GAN which generates images based on random noise. During image generation, we modulate deep features of input random noise from coarse-to-fine by injecting an initially restored image and the hole regions in multiple scales. We argue that during hole filling, the pixels near the hole boundary should be more deterministic (*i.e.*, with higher probability trusting the context and initially restored image to create natural inpainting boundary), while those pixels lie in the center of the hole should enjoy more degrees of freedom (*i.e.*, more likely to depend on the random noise for enhancing diversity). To this end, we propose spatially probabilistic diversity normalization (SPDNorm) inside the modulation to model the probability of generating a pixel conditioned on the context information. SPDNorm dynamically balances the realism and diversity inside the hole region, making the generated content more diverse towards the hole center and resemble neighboring image content more towards the hole boundary. Meanwhile, we propose a perceptual diversity loss to further empower PD-GAN for diverse content generation. Experiments on benchmark datasets including CelebA-HQ, Places2 and Paris Street View indicate that PD-GAN is effective for diverse and visually realistic image restoration.*

1. Introduction

There is a growing attention on developing advanced image inpainting methods for content removal [34, 3] and

*X. Han is the corresponding author. The results and code are available at https://github.com/KumapowerLIU/PD-GAN.

image restoration [39, 42, 45, 43]. Based on deep CNNs, image inpainting methods [31, 26, 54, 27] typically utilize an encoder-decoder network to generate meaningful image content for hole filling. Meanwhile, the content across the hole boundary is enforced consistent during visually realistic generation. By taking the input image with hole regions, the encoder captures deep image representations hierarchically, which are decoded to produce the output result.

While deep encoder-decoders improve the image inpainting performance, they target for single image generation that each input image corresponds to one restored result. In practice, image inpainting may produce multiple results because of the uncertain content generation within the hole region. This diverse image inpainting is less touched by existing inpainting methods. Recently, investigations [58, 57, 11] on diverse image inpainting follow the encoder-decoder structure. They are not effective to generate both diverse and realistic contents. One reason is that these methods still utilize the encoder to model the current masked image to Gaussian distribution and decode to a completed image, the variation of distribution is greatly limited by the masked image itself which leads to a decline in diversity, especially when the hole regions are free-form. On the other hand, these encoder-decoder networks utilize image reconstruction loss [18] during training. The generated content is thus enforced to be similar to the ground truth across both low-level and semantic representations. Heavily relying on such reconstruction loss limits the diverse content generation.

In this work, we propose PD-GAN, a diverse image inpainting network built upon a vanilla GAN. We notice that GAN is powerful to generate diverse image content based on different random noise inputs. Thus, instead of sending input images to the CNN, our PD-GAN starts from a random noise vector and then decodes this noise vector for content generation. In all the decoder layers, we inject prior information (coarse reconstruction result from a pre-trained partial convolution model [25]) and the region mask. The injection is fulfilled by the proposed SPDNorm (spatially probabilistic diversity normalization) module. SPDNorm gradually modulates deep features of the noise vector with input image representations. Specifically, the SPDNorm module learns a spatial transformation containing both hard and soft probabilistic diversity maps for feature fusion. The diversity is enhanced towards the hole center while is reduced towards the hole boundary.

Moreover, we propose a perceptual diversity loss to empower the diverse generation ability of PD-GAN. For two output images generated by the same prior information but input noise vectors, the perceptual diversity loss forces these two images to be farther in feature space. By training PD-GAN with the perceptual diversity loss, we can effectively generate both diverse and visually realistic contents for image inpainting. Some results can be found in Fig 1.

Our contributions are summarized as follows:

- Based on a vanilla GAN, the proposed PD-GAN modulates deep features of random noise vector via the proposed SPDNorm to incorporate context constraint.
- We propose a perceptual diversity loss to empower the network diversity.
- Experiments on the benchmark datasets indicate that our PD-GAN is effective to generate diverse and visually realistic contents for image inpainting.

2. Related Work

Image Inpainting. Existing inpainting methods can be divided into two categories: single-solution inpainting methods and diverse inpainting methods. The single-solution inpainting methods produce a single result for each masked image, while the diverse inpainting methods generate multiple results for each corrupted image.

Single-solution inpainting methods: Some traditional single-solution image inpainting methods [4, 21, 2] based on diffusion techniques propagate the contextual appearances to the missing regions. Other methods [5, 6, 37, 3, 50] based on patch match fill missing regions by calculating the statistics of patch offsets and transferring similar patches from the undamaged region to the hole region. The deep learning based single-solution inpainting methods [31, 16, 53, 29, 49, 32, 54, 36, 25, 51, 55, 27, 46, 26, 52, 9, 8, 12] typically involve the generative adversarial networks [10] to learn the semantic of image. These single-solution inpainting methods achieve high performance in predicting deterministic result for the hole regions, but they cannot generate a variety of semantically meaningful results.

Diverse inpainting methods: To generate pluralistic results given a corrupted image, [57, 58, 11] train conditional VAE [41] type of encoder-decoder networks. They encode the masked image to condition a Gaussian distribution, from which stochastic sampling at the test time achieves diverse inpainting results. However, the degree of diversity is controlled by the dispersion of distribution, and the dispersion is limited by the masked image. In contrast, our method samples a latent vector from the standard Gaussian distribution and map the latent vector to image directly by a single decoder. Meanwhile, these methods mainly rely on the image reconstruction loss during training, but unconditionally forcing the results similar to the ground truth will decrease the diversity of outputs. We propose the perceptual diversity loss to handle this issue.

Diverse image generation methods. In addition to diverse inpainting methods, the diverse image generation methods, such as VAE [19], GANs [10] and CVAE [41], can also generate the diverse results. BicycleGAN [60] makes invertible connections between the latent code and the generated im-

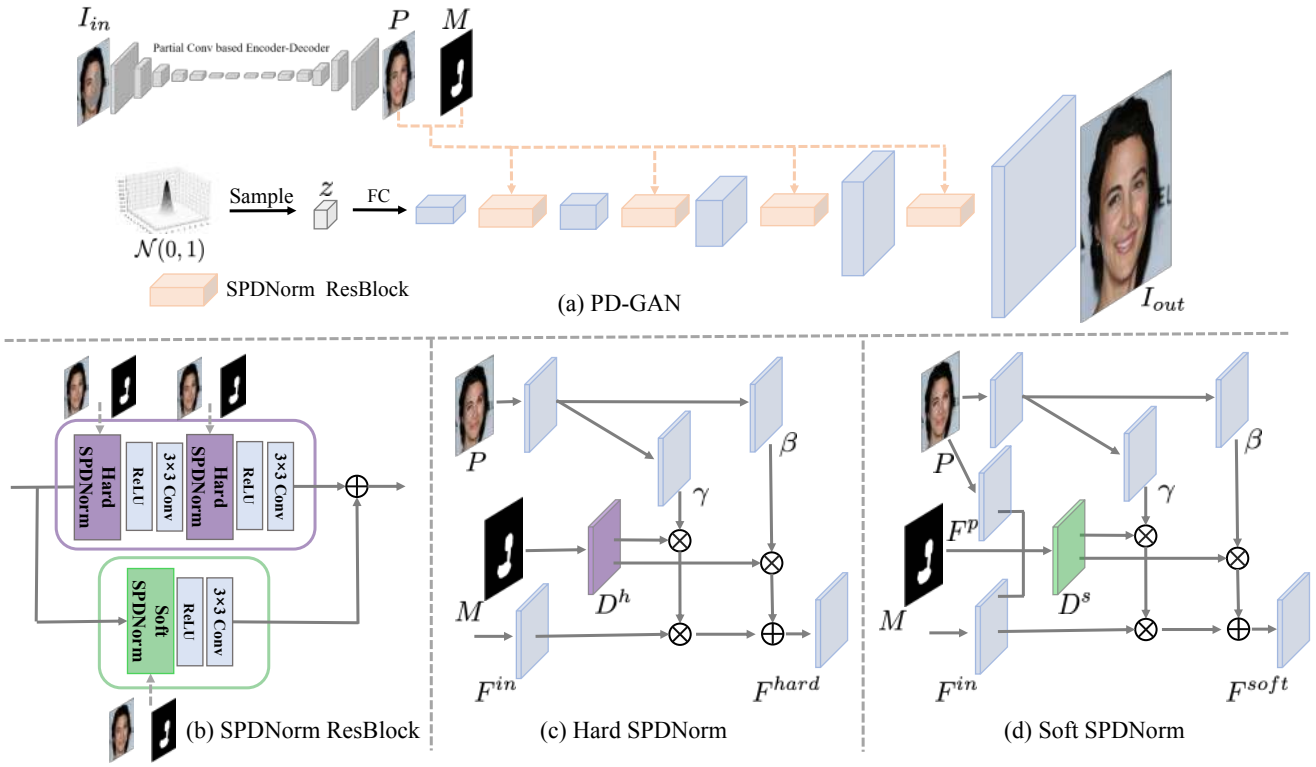


Figure 2. An overview of PD-GAN with SPDNorm. (a) We first get a coarse prediction as the prior information from a pre-trained partial conv [25] based network. Then we sample the latent vector z from a standard Gaussian distribution and PD-GAN modulates z with the prior information based on the SPDNorm Residual block. (b) The SPDNorm Residual block consists of hard and soft SPDNorm. (c-d) The hard and soft SPDNorm control the confidence of prior information based on the mask itself and a learning process, respectively.

age, which helps produce diverse results. MUNIT [15] and EBIT [56] combine the content and style from different images to achieve diverse image-to-image translation. However, these methods are not designed for image inpainting task, so the inpainted results usually present artifacts.

Normalization Layers. There are wide investigations of the normalization layers [14, 33, 17, 48, 1, 40, 38, 22] in deep learning to improve the network prediction performance. Among them, Spatially-adaptive denormalization (SPADE) [30] relates to our SPDNorm in that SPADE is a simple form of SPDNorm. By fully relying on the prior information, SPADE is not effective to empower network for diverse content output.

3. PD-GAN

Fig. 2 shows an overview of probabilistic diverse generative adversarial network (PD-GAN), which sets the coarse result as prior information and modulate the latent vector z to image space by a single decoder similar to a vanilla GAN. We utilize the pre-trained Partial Convolutional encoder-decoder [25] to get the coarse prediction. The coarse prediction and mask image are sent to SPDNorm Residual Blocks to provide prior knowledge for the generation process. The SPDNorm Residual Block consists of SPDNorm with hard and soft probabilistic diversity maps (Hard SPDNorm and Soft SPDNorm). The pixels close to the hole boundary should be more deterministic and the probability of generating diverse result is small, while those pixels at the center of the hole should enjoy more degrees of freedom and the probability of generating diverse result is large. The hard SPDNorm controls the probability according to the distance between the pixel and hole boundary, while the soft SPDNorm learns the probability in an adaptive process. In this section, we first describe SPDNorm in detail.

3.1. SPDNorm

The SPDNorm learns the scale and bias to transform the feature map. It contains hard and soft SPDNorm layers that are controlled by hard and soft probabilistic diversity maps, respectively. The hard probabilistic diversity map D^h is determined by the inpainting mask M without learning process. The soft probabilistic diversity map D^s is an adaptive map which is obtained by the input feature and coarse prediction with a learning process. As shown in Figure 2(c-d), We denote $F^{\text{in}} \in \mathbb{R}^{H\times W\times C}$ as the input feature map and

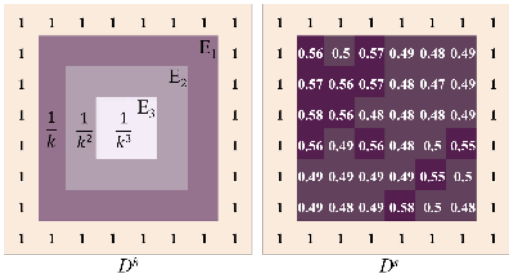


Figure 3. An example of hard probabilistic diversity map D^h and soft probabilistic diversity map D^s. The area with a value of 1 is the background. We show D^h and D^s at the stage with the feature map size of 8×8. The value of D^h gradually decreases from boundary to center. The value of D^s changed adaptively with a learning process.

$P \in \mathbb{R}^{H \times W \times 3}$ as the prior information (coarse prediction). We denote the outputs of hard and soft SPDNorm as F^{hard} and F^{soft}, respectively:

$$F^{hard}_{x,y,c} = D^h_{x,y}(\gamma_{x,y,c}(P)\frac{F^{\text{in}}_{x,y,c} - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} + \beta_{x,y,c}(P)), \quad (1)$$

$$F^{soft}_{x,y,c} = D^s_{x,y}(\gamma_{x,y,c}(P)\frac{F^{\text{in}}_{x,y,c} - \mu_c}{\sqrt{\sigma_c^2 + \epsilon}} + \beta_{x,y,c}(P)), \quad (2)$$

where $\gamma_{x,y,c}(P)$ and $\beta_{x,y,c}(P)$ are two variables output by two convolutional layers to control the element-wise influence from the coarse prior information P. $\mu_c = \frac{1}{H \times W}\sum_{x=1}^{H}\sum_{y=1}^{W} F^{\text{in}}_{x,y,c}$, $\sigma_c = \sqrt{\frac{1}{H \times W}\sum_{x=1}^{H}\sum_{y=1}^{W}\left(F^{\text{in}}_{x,y,c} - \mu_c\right)^2}$. Fig. 3 shows an example of D^h and D^s.

Hard SPDNorm. Intuitively, the closer the content is to the boundary, the stronger the constraints of the context. The content that is close the hole boundary needs to be consistent with the context, and thus we need more prior information to guide the hole filling process. While for the regions close to the hole center, PD-GAN needs less prior information and has a higher probability to generate diverse results. We control the probability by a hard probabilisitc diversity map D^h as shown in Fig. 3. For the mask M (0 for the missing region, and 1 for background), we apply n iterative dilation operations to it. The dilation is realized by a mask update process [25]. Specifically, we denote the mask update process as F_m. The mask after the i-th dilation operation is M_i, which is obtained by applying the update process on the mask from previous step (*i.e.*, $M_i = F_m(M_{i-1})$ with $M_0 = M$). Mathematically, the mask update process F_m can be expressed as:

$$M_i(x,y) = \begin{cases} 1 & \text{if } \sum_{(a,b)\in\mathcal{N}(x,y)} M_{i-1}(a,b) > 0 \\ 0 & \text{otherwise} \end{cases}, \quad (3)$$

where $\mathcal{N}(x,y)$ denotes the 3×3 neighborhood region centered at location (x,y). Finally, we fill in the i-th dilated region $E_i = M_i - M_{i-1}$ in D^h with value $\frac{1}{k^i}$ as shown in Fig. 3. We empirically set $k = 4$ in our paper.

As a result, we can get D^h whose value in the hole regions is exponentially decreases as the location is closer to the hole center, and is 1 outside hole regions (*i.e.*, background). As the value decreases, the credibility of prior information becomes lower, and the probability of diversity increases. There is no learning mechanism in the generation of D^h (*i.e.*, D^h is fixed). Our hard SPDNorm cannot only intuitively make the generated content and the background area coherent, but also ensure the diversity of the prediction.

Note that we set the number of mask updates to $n = 2, 2, 4, 4, 4$ for each stage from deep to shallow layers as the resolution of input mask increases.

Soft SPDNorm. The Hard SPDNorm explicitly controls the probability of generating diverse results according to the D^h. However, this probability should also be dynamic, which depends on the prior information and the inpainting mask. In other words, our network should be learnable and have the ability of paying attention to certain regions conditioned on the prior information and the inpainting mask. To this end, we propose soft SPDNorm to adaptively learn the probability of producing multiple content to achieve better inpainting results. The soft SPDNorm extracts the feature from both P and F^{in} to predict a soft probabilistic diversity map, guiding the diverse inpainting process. As Fig. 2(d) illustrates, we extract the feature map F^p from prior information P by convolutional layers, then F^p and input feature map F^{in} are concatenated to get D^s:

$$D^s = \sigma(\text{Conv}([F^p, F^{\text{in}}]) \cdot (1 - M) + M, \quad (4)$$

where σ is the sigmoid activation function and the elements corresponding to background in D^s are set to 1. In order to achieve stable training and generate plausible results, D^s adaptively changes probability of borrowing information from P. We find that the value of D^s in the foreground region changes smoothly and is close to 0.5, so only relying on D^s is unable to measure probability of predicting diverse results .

SPDNorm ResBlock. As discussed above, the hard SPDNorm increases the probability of getting diversity results but reduces the quality of the results. In contrast, the soft SPDNorm can stabilize the training and dynamically learn the condition of the prior information but lack diversity. So we propose the SPDNorm ResBlock to let them complement each other as shown in Fig. 2. Meanwhile, note that each residual block operates at a different scale, so we downsample the prior information and mask to match the corresponding spatial resolution.

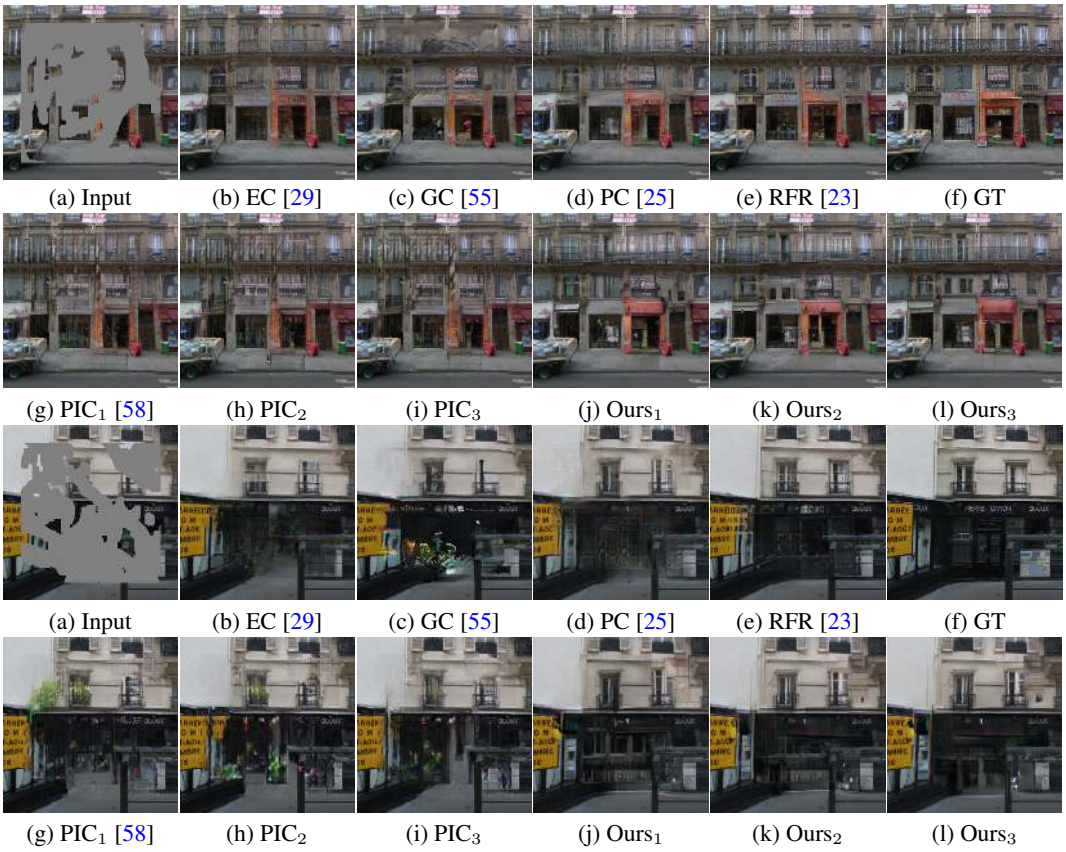


Figure 4. Qualitative comparisons with state-of-the-art methods on Paris Street View. Original images is in (f). Input images are in (a). The prior information is the output of PC in (d). The diverse outputs of PIC [58] are in (g)-(i). The diverse outputs of our method are in (j)-(l).

3.2. Perceptual Diversity Loss

We make use of the same set of reconstruction losses between the generated image and the ground truth as in [30]. However, only minimizing reconstruction loss inhibits the diversity of results, since it essentially forces the model to learn a deterministic single mapping. [28] proposes a diversity loss to activate the diversity of generative models, which could be utilized in PG-GAN. We denote the generator of PD-GAN as G, and $I_{out1} = G(z_1, P, M)$ and $I_{out2} = G(z_2, P, M)$ are two generated images conditioned on same coarse prediction P and image mask M but different latent vector z_1 and z_2. The diversity loss L_{div} proposed in [28] is as follows:

$$L_{div} = \frac{\|z_1 - z_2\|_1}{\|I_{out1} - I_{out2}\|_1 + \varepsilon}, \tag{5}$$

which forces two output images to be farther in pixel space if their corresponding latent codes are far from each other. However, we find L_{div} is not suitable for the diverse inpainting task. First of all, minimizing L_{div} changes the content of the contextual regions that should be constant for different latent vectors. Moreover, we find the training very unstable. Minimizing L_{div} promotes the results to be all black or all white in order to maximize the pixel distance between two output images. In this paper, we propose a simple but effective perceptual diversity loss L_{pdiv} that tackles the above issues:

$$L_{pdiv} = \frac{1}{\sum_i \|F_i(I_{out1}) \cdot M - F_i(I_{out2}) \cdot M\|_1 + \varepsilon}, \tag{6}$$

where F_i is the i-th layer of a VGG-19 network [35] pretrained on ImageNet. In our work, F_i corresponds to the activation maps from layers ReLU1_1, ReLU2_1, ReLU3_1, ReLU4_1, and ReLU5_1. L_{pdiv} keeps the context unchanged by introducing the mask in the loss. Meanwhile, the L_{pdiv} is calculated on the perceptual space instead of raw pixel space. Maximizing the distance in highly nonlinear network feature space integrates semantic measurement and avoids trivial solution that completely generates black or white pixels. Note that we do not involve the latent vectors in L_{pdiv} as we want to maximize the perceptual distance of generated images no matter how close their latent vectors are. We find this further stabilizes the training.

In addition to the perceptual diversity loss, we follow the SPADE [30] and utilize the reconstruction loss [18], feature matching loss [44] and hinge adversarial loss [24] to opti-

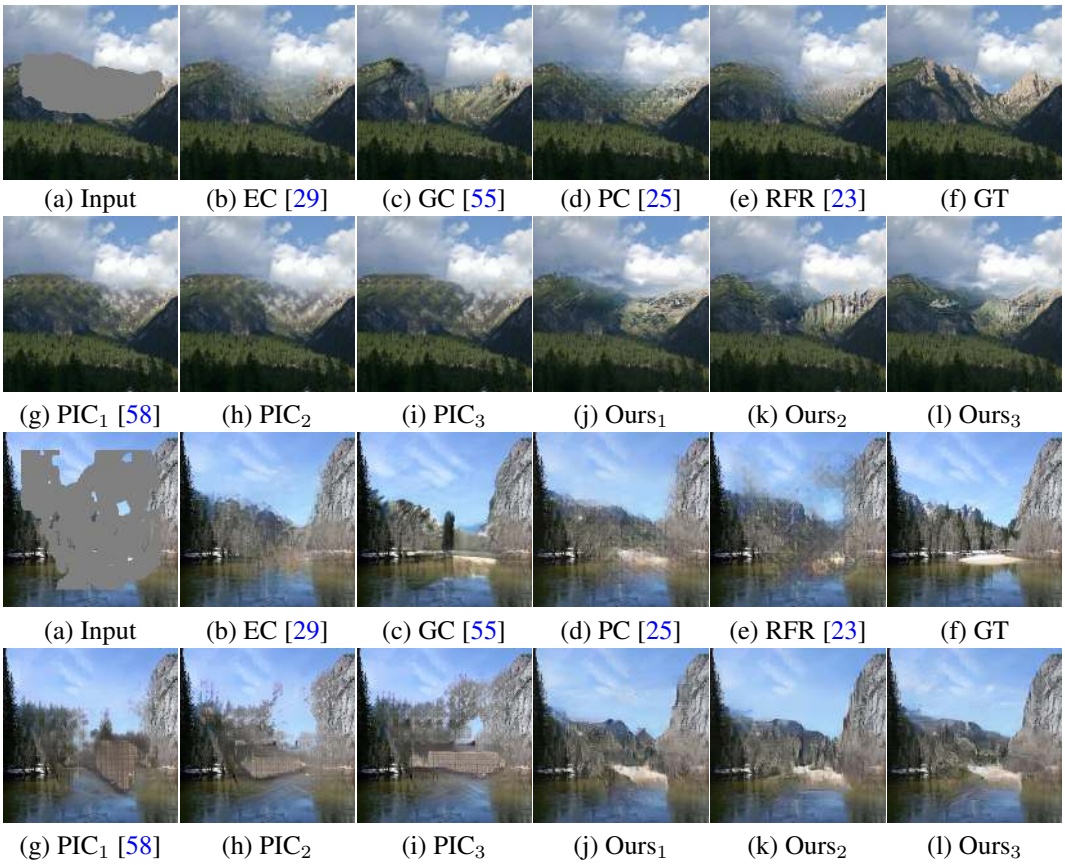


Figure 5. Qualitative comparisons with state-of-the-art methods on Place2. Original images is in (f). Input images are in (a). The prior information is the output of PC in (d).The diverse outputs of PIC [58] are in (g)-(i). The diverse outputs of our method are in (j)-(l).

mize our network.

4. Experiments

Datasets. We evaluated our proposed model on three datasets including Paris StreetView [7], CelebA-HQ [20], and Places2 [59]. For the Paris StreetView [7] and CelebA-HQ [20], we use their original training and test splits. For the Places2 [59], we select the butte, canyon, field, moutain, mountain-path, mountain-snowy, sky, tundra and valley scene categories to train and validate our model following the shift-net [51]. Since our model can generate multiple outputs, we randomly sampled 100 images for each masked image, and chose the top 5 results based on the discriminator scores for evaluation.

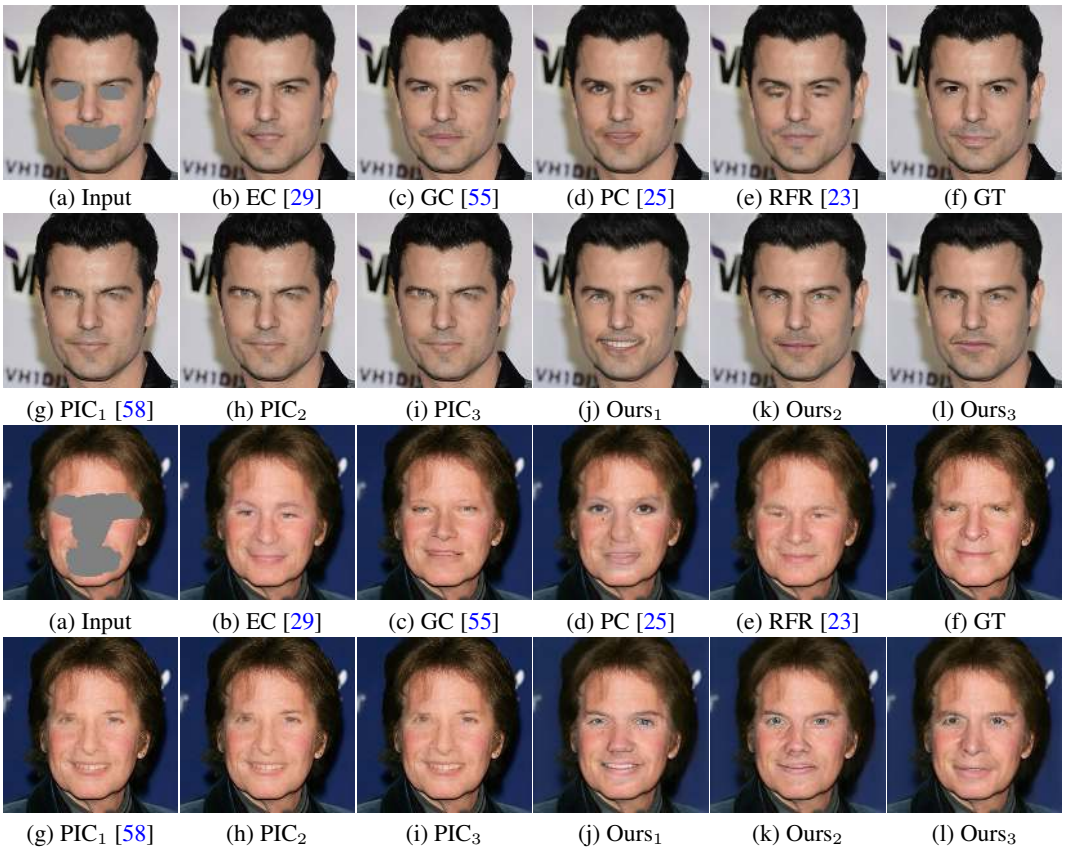
Compared Methods. We compare with the following inpainting approaches: recurrent feature reasoning (RFR) [23], partial conv (PC) [25], gated conv (GC) [55], edge connect (EC) [29], and PICNet (PIC) [58]. Plus, we compare with CVAE [41] and BicycleGAN [60] on the ability to generate diverse results.

Implementation Details. All of our models are trained on irregular masks [25]. The mask and image are resized to 256×256 as network input. We train the PC [25] following the official implementation. We set the inpainting results of the pre-trained PC [25] as the prior information. Our model is optimized using Adam optimizer with $\beta_1 = 0.0$ and $\beta_2 = 0.99$. The initial learning rate is 1×10^{-4}, and we utilize the TTUR [13] strategy to train our model. We train the network for 500K iterations with batch size of 6. We choose low-dimensional manifold vector $|z| = 128$ across all the datasets following SPADE [30].

4.1. Comparison with Existing Work

Qualitative Comparisons. The qualitative comparisons on the results for filling irregular holes on Paris StreetView [7], CeleBA-HQ [20] and Place2 [59] are shown in Fig 4, Fig 6 and Fig 5 respectively. The result of PC in (d) is used as the prior information of our model.

For the Paris StreetView, although EC, PC and RFR can generate roughly correct structure, the results still contain blurry and unrealistic textures as shown in (b), (d) and (e), respectively. PIC is not effective to recover image content, and the results lack diversity. For CeleBA-HQ, the single-



(a) Input (b) EC [29] (c) GC [55] (d) PC [25] (e) RFR [23] (f) GT

(g) PIC_1 [58] (h) PIC_2 (i) PIC_3 (j) $Ours_1$ (k) $Ours_2$ (l) $Ours_3$

(a) Input (b) EC [29] (c) GC [55] (d) PC [25] (e) RFR [23] (f) GT

(g) PIC_1 [58] (h) PIC_2 (i) PIC_3 (j) $Ours_1$ (k) $Ours_2$ (l) $Ours_3$

Figure 6. Qualitative comparisons with state-of-the-art methods on CelebA-HQ. Original images is in (f). Input images are in (a). The prior information is the output of PC in (d).The diverse outputs of PIC [58] are in (g)-(i). The diverse outputs of our method are in (j)-(l).

Table 1. Numerical comparisons on the Place2 dataset. ↓ indicates lower is better while ↑ indicates higher is better.

	PSNR ↑				SSIM ↑				FID ↓			
Mask	10-20%	20-30%	30-40%	40-50%	10-20%	20-30%	30-40%	40-50%	10-20%	20-30%	30-40%	40-50%
EC	28.82	26.19	24.64	**23.29**	0.929	0.877	0.833	**0.783**	21.86	37.87	48.06	59.27
GC	28.64	25.98	24.48	22.96	0.928	0.879	0.839	0.784	20.29	35.71	45.16	56.07
PC	28.34	25.83	24.29	22.58	0.920	0.862	0.815	0.772	21.94	37.32	47.59	59.29
RFR	29.16	26.38	24.19	22.43	0.923	0.859	0.774	0.681	21.62	36.72	59.00	85.64
PIC	28.38	25.66	23.92	22.46	0.921	0.860	0.807	0.744	38.64	58.83	69.46	99.17
Ours	**29.20**	**26.75**	**25.48**	23.15	**0.935**	**0.880**	**0.839**	0.782	**19.98**	**34.84**	**44.24**	**52.68**

solution inpainting methods can generate natural but blurry content. The generated content of diverse image inpainting method PIC is basically the same and seems blurry. For Place2, EC, PC and RFR get blurry and unnatural predictions. Although more visually pleasing content can be generated by GC as shown in (c), GC cannot produce diverse results. PIC can generate different content for a single masked input, however it lacks of obvious differences while unreasonable semantics are rendered in the filling regions. In contrast to the above methods, our model generates multiple results with higher naturalness.

Quantitative Comparisons. Since our model is used to solve the diverse image inpainting task, the generated results need to have both authenticity and diversity. We compare our method with baselines from two aspects: Realism and Diversity. For realism comparison, our model is compared with both single- and diverse-solution inpainting methods on Place2. We follow PIC and assume that one of our top 5 samples (ranked by the discriminator) will be close to the original ground truth, and select the single sample with the best balance of quantitative measures for comparison. For the evaluation metrics, we use the SSIM [47], PSNR and FID [13]. The evaluation results are shown in Table 1. For each hole versus image ratio, we randomly select 10 masks for testing. Our method outperforms existing

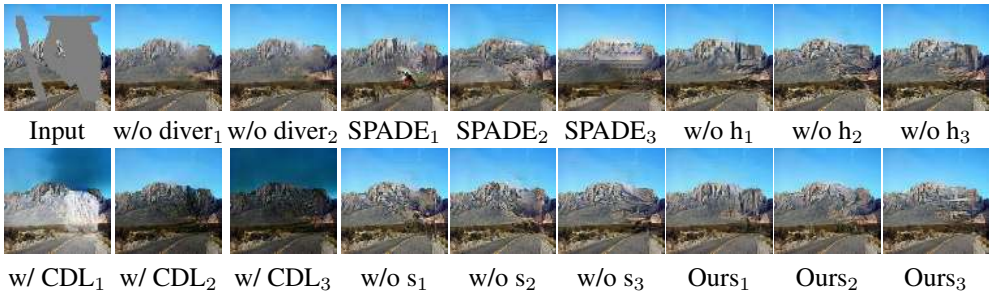


Figure 7. Visual results of ablation study. The nodiver$_{1\sim2}$ are the diverse results of our methods without diversity loss. The w/o h$_{1\sim3}$ are results after we replace the hard SPDNorm with soft. The w/o s$_{1\sim3}$ are results after we replace the soft SPDNorm with hard. The predictions of our methods with conventional diversity loss [28] are shown in CDL$_{1\sim3}$. The results of SPADE [30] are in SPADE$_{1\sim3}$. Our results are in Ours$_{1\sim3}$.

Table 2. Quantitative comparison of diversity with the state of-the-art methods on the Places2 dataset.

	Diversity (LPIPS)↑			
Method	CVAE	BicycleGAN	PIC	Ours
I_{out}	0.0432	0.0925	0.1096	**0.1238**
$I_{out(m)}$	0.0506	0.1032	0.1547	**0.1799**

Table 3. Quantitative ablation study on the Places2 dataset. ↓ indicates lower is better while ↑ indicates higher is better.

	PSNR ↑	SSIM ↑	FID ↓	LIPIS ↑
w/ CDL	22.48	0.732	61.55	0.198
w/o diver	**23.63**	**0.796**	56.55	0.1934
w/o h	23.03	0.771	61.22	0.2193
w/o s	22.48	0.762	61.48	**0.2215**
SPADE	22.26	0.779	60.55	0.198
Ours	23.15	0.782	**52.68**	0.2206

methods to fill irregular holes under various situations.

For the diversity comparison, we utilize the LPIPS metric [60, 58] to calculate the diversity score. The average score is calculated between 5K pairs generated from a sampling of 1K images of Places2 dataset. I_{out} and $I_{out(m)}$ are the full image output and mask-region output, respectively. Our method obtains relatively higher diversity scores than other existing methods as shown in Table 2. To further demonstrate the superior of PD-GAN, we conduct extra user studies with 10 volunteers. Each subject is asked to compare 10 sets of inpainting results of PD-GAN and PIC and select the method with diverse. PD-GAN is favored in **83%** of cases.

5. Ablation Study

SPDNorm. To evaluate the effects of SPDNorm, we compare the following ablations: 1) Replacing all the soft SPDNorm with hard SPDNorm (w/o s); 2) replacing all the hard SPDNorm with soft SPDNorm (w/o h); 3) Replacing all the SPDNorm with SPADE [30], which is equivalent to setting D_h and D_s to 1. Thus, SPADE can be regarded as a degenerated form of the proposed SPDNorm. As shown in Fig 7, SPADE makes training unstable and generates worse results, since SPADE unconditionally relies on the coarse prior information, which is contrary to the purpose of generating diverse results. The outputs of our method without soft SPDNorm (w/o s) has diverse details, but the artifacts are obvious. In comparison, the predictions of our method without hard SPDNorm (w/o h) contain meaningful content, but the diversity is declined. By utilizing both soft and hard SPDNorm, our method achieves favorable results on both diversity and quality. Table 3 shows the similar numerical performance on the Places2 dataset where the combination of hard and soft SPDNorm is suitable for the diverse image inpainting task. We choose the mask ratio $40-50\%$ here.

Perceptual Diversity Loss. We show the contributions of perceptual diversity loss (Eqn. 6) by removing it (w/o diver) or replacing it with conventional diversity loss [28] in Eqn. 5 (w/ CDL). As shown in Fig 7, without using perceptual diversity loss the content generated lacks diversity. The content generated with the constraint of conventional diversity loss are more diverse. However, the recovered content tend to be all black or all white. The proposed perceptual diversity loss can solve the above issues. Similar performance has been shown numerically in Table 3 where our full method achieves favorable results.

6. Conclusion

We propose a novel probabilistic diverse GAN (PD-GAN) for image inpainting. To get diverse inpainting results, PD-GAN utilizes prior information to modulate a random noise progressively. For the modulation process, PD-GAN adopts both soft and hard spatially probabilistic diversity normalization (SPDNorm) to control the probability of producing diverse results. Meanwhile, we propose the perceptual diversity loss to further boost diversity of PD-GAN. Experiments on a variety of datasets demonstrate that our PD-GAN cannot only produce diverse prediction, but also generates high-quality reconstruction content.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[2] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *TIP*, 2001.

[3] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan Goldman. Patchmatch: A randomized correspondence algorithm forstructural image editing. In *SIG*, 2009.

[4] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *SIG*, 2000.

[5] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *TIP*, 2004.

[6] Soheil Darabi, Eli Shechtman, Connelly Barnes, Dan B Goldman, and Pradeep Sen. Image melding: combining inconsistent images using patch-based synthesis. *TOG*, 2012.

[7] Carl Doersch, Saurabh Singh, Abhinav Gupta, Josef Sivic, and Alexei A Efros. What makes paris look like paris? *TOG*, 2015.

[8] Haoye Dong, Xiaodan Liang, Xiaohui Shen, Bochao Wang, Hanjiang Lai, Jia Zhu, Zhiting Hu, and Jian Yin. Towards multi-pose guided virtual try-on network. In *ICCV*, October 2019.

[9] Haoye Dong, Xiaodan Liang, Yixuan Zhang, Xujie Zhang, Xiaohui Shen, Zhenyu Xie, Bowen Wu, and Jian Yin. Fashion editing with adversarial parsing learning. In *CVPR*, pages 8120–8128, 2020.

[10] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *NIPS*, 2014.

[11] Xintong Han, Zuxuan Wu, Weilin Huang, Matthew R Scott, and Larry S Davis. Finet: Compatible and diverse fashion image inpainting. In *ICCV*, pages 4481–4491, 2019.

[12] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S. Davis. Viton: An image-based virtual try-on network. In *CVPR*, June 2018.

[13] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *NIPS*, 2017.

[14] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *ICCV*, pages 1501–1510, 2017.

[15] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, pages 172–189, 2018.

[16] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Globally and locally consistent image completion. In *SIG*, 2017.

[17] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[18] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, 2016.

[19] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[20] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *CVPR*, pages 5549–5558, 2020.

[21] Anat Levin, Assaf Zomet, and Yair Weiss. Learning how to inpaint from global image statistics. In *ICCV*, 2003.

[22] Boyi Li, Felix Wu, Kilian Q Weinberger, and Serge Belongie. Positional normalization. In *NIPS*, pages 1622–1634, 2019.

[23] Jingyuan Li, Ning Wang, Lefei Zhang, Bo Du, and Dacheng Tao. Recurrent feature reasoning for image inpainting. In *CVPR*, June 2020.

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

[25] 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 *ECCV*, 2018.

[26] 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.

[27] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *ICCV*, 2019.

[28] Qi Mao, Hsin-Ying Lee, Hung-Yu Tseng, Siwei Ma, and Ming-Hsuan Yang. Mode seeking generative adversarial networks for diverse image synthesis. In *CVPR*, pages 1429–1437, 2019.

[29] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. In *International Workshop on Applications of Computer Vision Workshops*, 2019.

[30] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *CVPR*, pages 2337–2346, 2019.

[31] Deepak Pathak, Philipp Krahenbuhl, Jeff Donahue, Trevor Darrell, and Alexei Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

[32] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *ICCV*, 2019.

[33] Tim Salimans and Durk P Kingma. Weight normalization: A simple reparameterization to accelerate training of deep neural networks. In *NIPS*, pages 901–909, 2016.

[34] Rakshith R Shetty, Mario Fritz, and Bernt Schiele. Adversarial scene editing: Automatic object removal from weak supervision. In *NIPS*, pages 7706–7716, 2018.

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

[36] Chao Song, Yuhangand Yang, Zhe Lin, Xiaofeng Liu, Qin Huang, Hao Li, and CC Jay. Contextual-based image inpainting: Infer, match, and translate. In *ECCV*, 2018.

[37] Yibing Song, Linchao Bao, Shengfeng He, Qingxiong Yang, and Ming-Hsuan Yang. Stylizing face images via multiple exemplars. *CVIU*, 2017.

[38] Yibing Song, Chao Ma, Lijun Gong, Jiawei Zhang, Rynson WH Lau, and Ming-Hsuan Yang. Crest: Convolutional residual learning for visual tracking. In *ICCV*, 2017.

[39] Yibing Song, Jiawei Zhang, Lijun Gong, Shengfeng He, Linchao Bao, Jinshan Pan, Qingxiong Yang, and Ming-Hsuan Yang. Joint face hallucination and deblurring via structure generation and detail enhancement. *IJCV*, 2019.

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

[41] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *ECCV*, pages 835–851. Springer, 2016.

[42] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *CVPR*, pages 2747–2757, 2020.

[43] Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Old photo restoration via deep latent space translation. *arXiv preprint arXiv:2009.07047*, 2020.

[44] 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 *CVPR*, pages 8798–8807, 2018.

[45] Yinglong Wang, Yibing Song, Chao Ma, and Bing Zeng. Rethinking image deraining via rain streaks and vapors. In *ECCV*, 2020.

[46] Yi Wang, Xin Tao, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Image inpainting via generative multi-column convolutional neural networks. In *NIPS*, 2018.

[47] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *TIP*, 2004.

[48] Yuxin Wu and Kaiming He. Group normalization. In *ECCV*, pages 3–19, 2018.

[49] Wei Xiong, Jiahui Yu, Zhe Lin, Jimei Yang, Xin Lu, Connelly Barnes, and Jiebo Luo. Foreground-aware image inpainting. In *CVPR*, 2019.

[50] Zongben Xu and Jian Sun. Image inpainting by patch propagation using patch sparsity. *TIP*, 2010.

[51] Zhaoyi Yan, Xiaoming Li, Mu Li, Wangmeng Zuo, and Shiguang Shan. Shift-net: Image inpainting via deep feature rearrangement. In *ECCV*, 2018.

[52] Zili Yi, Qiang Tang, Shekoofeh Azizi, Daesik Jang, and Zhan Xu. Contextual residual aggregation for ultra high-resolution image inpainting. In *CVPR*, June 2020.

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

[54] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *CVPR*, 2018.

[55] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Free-form image inpainting with gated convolution. In *ICCV*, 2019.

[56] Pan Zhang, Bo Zhang, Dong Chen, Lu Yuan, and Fang Wen. Cross-domain correspondence learning for exemplar-based image translation. In *CVPR*, June 2020.

[57] Lei Zhao, Qihang Mo, Sihuan Lin, Zhizhong Wang, Zhiwen Zuo, Haibo Chen, Wei Xing, and Dongming Lu. Uctgan: Diverse image inpainting based on unsupervised cross-space translation. In *CVPR*, pages 5741–5750, 2020.

[58] Chuanxia Zheng, Tat-Jen Cham, and Jianfei Cai. Pluralistic image completion. In *CVPR*, pages 1438–1447, 2019.

[59] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *PAMI*, 2017.

[60] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. In *NIPS*, pages 465–476, 2017.