This CVPR workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Internal Diverse Image Completion

Noa Alkobi
Technion

Tamar Rott Shaham
Technion, MIT

Tomer Michaeli
Technion

Abstract

Image completion is widely used in photo restoration and editing applications, e.g. *for object removal. Recently, there has been a surge of research on generating diverse completions for missing regions. However, existing methods require large training sets from a specific domain of interest, and often fail on general-content images. In this paper, we propose a diverse completion method that does not require a training set and can thus treat arbitrary images from any domain. Our internal diverse completion (IDC) approach draws inspiration from recent single-image generative models that are trained on multiple scales of a single image, adapting them to the extreme setting in which only a small portion of the image is available for training. We illustrate the strength of IDC on several datasets, using both user studies and quantitative comparisons.*

1. Introduction

Image completion (or inpainting) refers to the problem of filling-in a missing region within an image. This task is of wide applicability in image restoration, editing and manipulation applications (*e.g.* for object removal), and has thus seen intense research efforts over the years. Since the first inpainting method of Bertalmio et al. [4], image completion techniques have rapidly evolved from methods that locally diffuse information into the missing region [1, 5, 38], to algorithms that copy chunks from other locations within the image [2, 30, 43], and recently, to deep learning approaches that are trained on large external datasets [28, 31, 41, 47].

The transition to externally trained models has enabled the treatment of challenging inpainting settings, like completion of large parts of semantic objects, and *diverse inpainting* [6, 29, 51, 53], *i.e.* generating a variety of different completions. However, despite their merits, externally trained methods are limited by the need for very large training sets from the particular domain of interest (*e.g.* faces [16, 24], bedrooms [46], etc.). Even when trained on more generic datasets, like Places [54] or Imagenet [9], such models are still often limited in the object categories or photograph styles they can handle. An additional limitation of external models is that they cannot be trained on the task of object removal (as opposed to generic inpainting), because this requires pairs of images, one depicting an object over some background, and one depicting only the background behind the object. Such pairs are obviously unavailable in real-world scenarios, but seem to be crucial for successful object removal. Indeed, when generic external inpainting methods are used for object removal, they tend to insert alternative objects instead of filling in the missing region with background. This is illustrated in Fig. 1. In light of those limitations, it is natural to ask whether external training is the only route towards generic diverse completion.

In this paper, we explore the idea of *internal diverse completion* (IDC). We propose an inpainting method that is capable of generating a plethora of different completions, without being exposed to any external training example [1]. IDC can be applied to images from any domain and is not constrained to particular image dimensions, aspect ratios, or mask shapes and locations. The reliance on internal learning, obviously limits IDC in its ability to complete large parts of semantic objects (*e.g.* parts of a face or an animal). However, in object removal scenarios, this is typically not required. We illustrate that in such cases, IDC performs at least on par with externally trained models when applied to images from the domain on which they were trained, while it can also be successfully applied to other domains. One representative within-domain example is shown in Fig. 1, where the external DSI [29], ICT [40], and CoMod-GAN [52] models do not provide plausible completions. This is while IDC manages to generate diverse photo-realistic completions for the grass. Completion examples for other domains are shown in Fig. 2.

We make use of recent progress in single-image generative models. Particularly, we adopt the framework of SinGAN [33], which is a multi-scale patch-GAN model that can be trained on a single natural image. However, our problem is more challenging than that treated in [33] because (i) only a small part of the image is available for training, especially at the coarser scales, and (ii) the generated

[1] Code is available at: https://github.com/NoaAlkobi/IDC

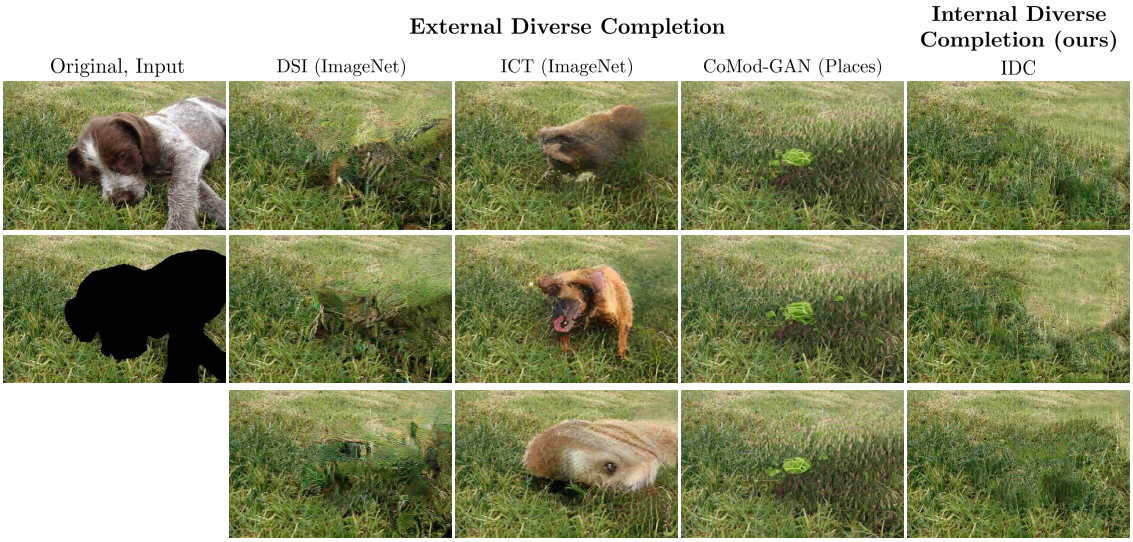


Figure 1. **Internal Diverse Completion.** We introduce a new single image GAN model targeted for the task of image completion. Our model is trained on a single image with a missing region, and then suggests diverse completions based on only the internal statistics of the image to be inpainted itself. In the context of object removal, externally trained models are often unsuitable, as they attempt to insert alternative objects in the missing region (similar faliures occure with DSI and ICT models trained on Places; see supplementary).

content must match and seamlessly blend with the neighboring regions. We address these difficulties through a dedicated architecture, and training and inference processes.

2. Related work

Internal inpainting Internal completion methods rely solely on the information within the image to be inpainted. Methods in this category include diffusion-based approaches [1, 3, 17, 38] and patch-based algorithms [2, 5, 8, 10, 30, 35, 43]. Recently, it was shown that internal inpainting can also be achieved by overfitting a deep neural network (DNN) to the input image [39]. However, despite the inherent ambiguity in the inpainting task, all internal methods to date are designed to output only a single completion. Here, we introduce an internal completion method that can generate diverse completion suggestions.

External inpainting A dramatic leap in the ability to complete semantic contents, was achieved by transitioning to DNNs trained on large datasets. Example approaches include GAN based techniques [14, 19, 45] and encoder-decoder frameworks [21, 22, 28]. Progress in this route has been driven by the developments of various mechanisms, including gated, partial and fast Fourier convolutions [20, 36, 48], multi-scale architectures [31, 44], edge map representations [27, 32, 42] and contextual attention [47]. All these methods require large training sets, and like their internal counterparts, output a single solution. By contrast, our method is trained only on the masked input image and generates diverse completions.

Diverse external inpainting Several recent works developed diverse inpainting methods based on VAEs [29,51,53], GANs [7, 23, 52], transformers [40, 49], and diffusion models [34]. These methods work well on inputs resembling their training data (in content and mask shapes). However, their performance degrades on out-of-distribution inputs. Our method does not require a training set and is thus not restricted to specific types of images or masks.

3. Method

We are given a masked image

$$y = x \odot (1 - m), \tag{1}$$

where x is the original (non-masked) image, m is a binary mask containing 1's in the masked regions and 0's elsewhere, and $\odot$ denotes per-pixel product. Our goal is to randomly generate estimates $\tilde{x}$ of x that satisfy two properties: (i) each $\tilde{x}$ matches x in the non-masked regions, and (ii) the distribution of patches within $\tilde{x}$ is similar to that of the non-masked patches of y.

Our approach is inspired by SinGAN [33], which is a multi-scale patch-GAN architecture that can be trained on a single image. Specifically, our goal is to train a model that captures the distribution of small patches within the valid regions of y, at multiple scales, and then use this model to generate content within the masked region. To this end, we construct pyramids $\{y_0, \ldots, y_N\}$ and $\{m_0, \ldots, m_N\}$, from

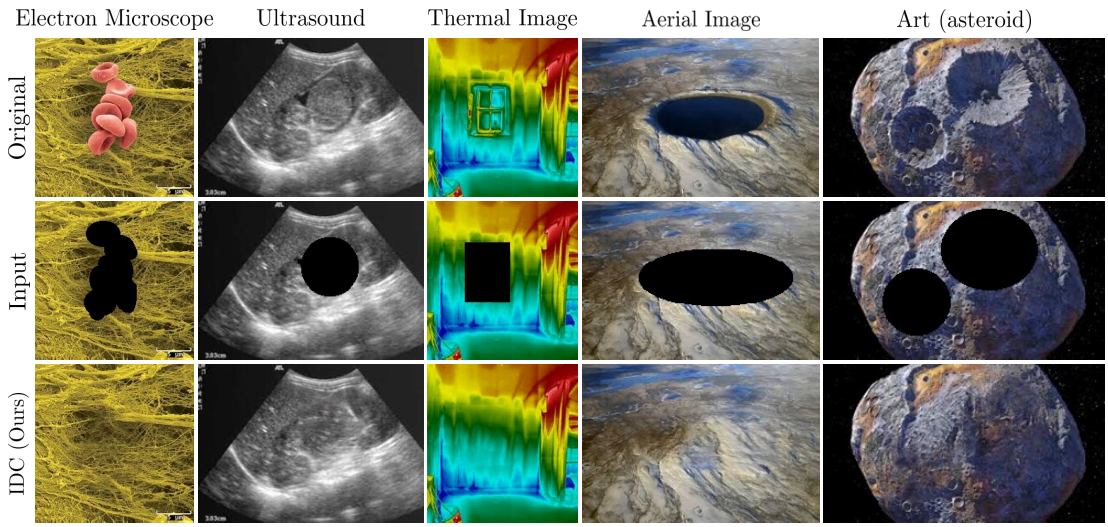


Figure 2. **Internal completion for different domains.** Our method can be applied to images from arbitrary domains and is not restricted to a specific mask shape or location. From left to right: blood cells imaged by an electron microscope, ultrasound image of a Wilms tumor, thermal image of a room, aerial photograph of Northern Quebec's Pingualuit Crater, and an artist's conception of the asteroid 16 Psyche. Completions with baselines are provided in SM.

the input image y and the binary mask m, where $y_0 = y$, $m_0 = m$, and the scale factor between consecutive levels is α. Using these images, we train a pyramid of GANs, as shown in Fig. 3. Specifically, in each pyramid level n, we train a generator G_n against a patch discriminator D_n [15, 18]. The generator consumes a white Gaussian noise map, z_n, having the same dimensions as the input image at that scale, as well as an up-sampled version of the fake image generated by the previous scale (the coarsest scale accepts only noise), and it outputs a fake image $\tilde{x}_n$. The discriminator D_n is presented with either the fake image $\tilde{x}_n$ or the masked image y_n (or a processed version of y_n, see Sec. 3.1), and attempts to classify each of the overlapping (valid) patches of its input as real or fake. Thus, its output is a discrimination map.

All the generators and discriminators comprise five convolutional blocks consisting of `conv-BN-LeakyReLU`. Therefore, each GAN in the pyramid captures the distribution of 11×11 patches (the networks' receptive field). At the coarser scales, these patches cover large structures relative to the image dimensions, whereas the finer scales capture smaller structures and textures.

3.1. Training

Training is done sequentially from the coarsest scale to the finest one. Each scale is trained individually, while keeping the generators of all coarser scales fixed. We optimize an objective function comprising an adversarial loss and a reconstruction loss.

For the *adversarial term*, we use the Wasserstein GAN loss, together with gradient penalty [11]. For this loss, the discrimination score for the masked image y_n is computed by masking the elements of the discrimination map that correspond to patches containing invalid (masked) pixels, while computing the BN statistics in each layer only over features that are not affected by invalid input pixels (see SM for the importance of BN masking). Note that when computing the discrimination score for a fake image $\tilde{x}$, there are no invalid pixels to ignore.

The goal of the *reconstruction loss* is to guarantee that there is a fixed input noise z_n^{rec}, which is mapped to an image that equals y_n at the valid regions. Therefore, we take this loss to be the MSE between $G_n(z_n^{\text{rec}}) \odot (1 - \tilde{m}_n)$ and $y_n \odot (1 - \tilde{m}_n)$, where $\tilde{m}_n$ is a soft version of m_n, whose boundaries gradually transition from 0 to 1. The reconstruction loss stabilizes training, and is also crucial for generating our completions at inference time (see Sec. 3.2 below).

Ideally, we would want all GANs in the pyramid to directly learn only from the valid patches in y_n (those not containing masked pixels), as described above. However, when the masked region is large, there often do not exist sufficiently many valid patches for stably training a GAN, especially at the coarse scales. This is illustrated in Fig. 4, where only patches whose center pixel is outside the black region, can be used for training. We, therefore, treat the coarse scales differently from the fine ones.

Coarse scales Let i be the finest scale in which less than 40% of the patches are valid. In all scales belonging to $\{i, \ldots, N\}$, we do not perform masking in the discrimina-

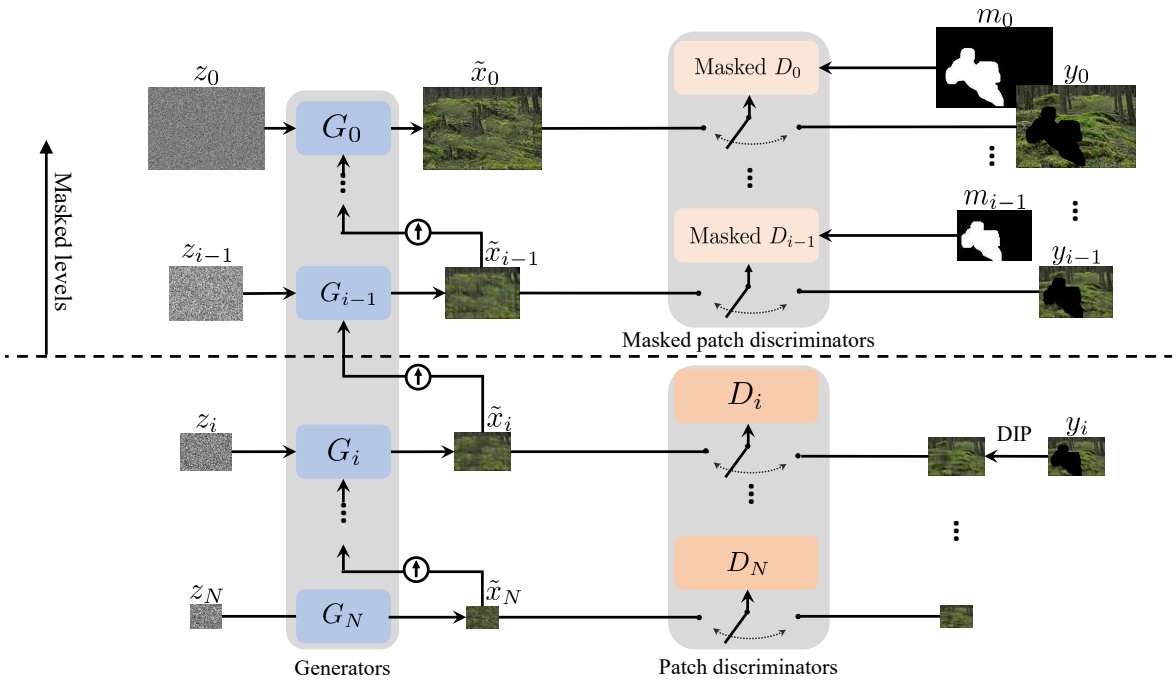


Figure 3. **Architecture.** Our model is composed of a pyramid of GANs, which we train sequentially from coarse to fine. Each generator consumes a noise map and the upsampled version of the image generated by the preceding scale, and it outputs a fake image. Each discriminator learns to distinguish generated images from the real one. At coarse scales, the real image presented to the discriminator is a naively inpainted version of the masked image. At fine scales, the real image is the masked image itself, and the discriminator ignores the missing region.

tor. Instead, we use a naively inpainted version of the input as our real image, which we obtain using an internal technique. This naive inpainting solution is generated at scale i and down-sampled to yield the real image at all coarser scales (see Fig. 3). Specifically, to obtain a naive completion, we adopt the deep image prior (DIP) method [39], and modify it to ensure color consistency (see SM for the importance of the modification). Specifically, we train a light-weight U-Net model that maps a fixed input (random noise) into an inpainted image. This is done by minimizing two loss terms: (i) MSE between the image y_i and the U-Net's output over the valid pixels, and (ii) MSE between each pixel in the inpainted region, and its nearest neighbor (in terms of RGB values) among the valid pixels in y_i. The latter term serves as an approximation to the KL divergence between the distributions of RGB values in the inpainted pixels and in the valid pixels [25]. Finally, we stitch the naively inpainted region with the valid image regions. Examples of naively inpainted solution and ablation of the method components are provided in the SM.

3.2. Inference

Once the model is trained, our goal is to generate only the missing region. Thus, instead of sampling full noise maps (as done during training), we construct noise maps containing z_n^{rec} at all elements affecting the valid regions, and new noise samples z_n at the rest of the elements. Note that each element of the noise map affects a region of the size of the receptive field in the generated output. Therefore, the new noise samples are generated only within an eroded version of the mask (by half a receptive field), so that the noise map is constructed as $z_n^{\text{test}} = z_n^{\text{rec}} \odot (1 - \text{erode}\{m_n\}) + z_n \odot \text{erode}\{m_n\}$. This is illustrated in Fig. 5. The final completion result is achieved by merging the generated inpainted region with the input image outside the mask. Since we want the fusion to be smooth, for this stage we use a soft version of the mask, obtained by convolving it with a Gaussian kernel with $\sigma = 5$.

4. Experiments

We now evaluate the performance of our method, qualitatively and quantitatively, and compare it to existing baselines. We focus on two scenarios: *object removal* and *generic inpainting*. In the former, which is commonly encountered in real applications, the goal is to edit an image so as to cut out a whole semantic object. In the latter, which is the focus of many papers, the goal is to treat arbitrary masks

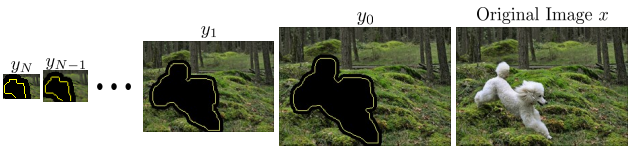
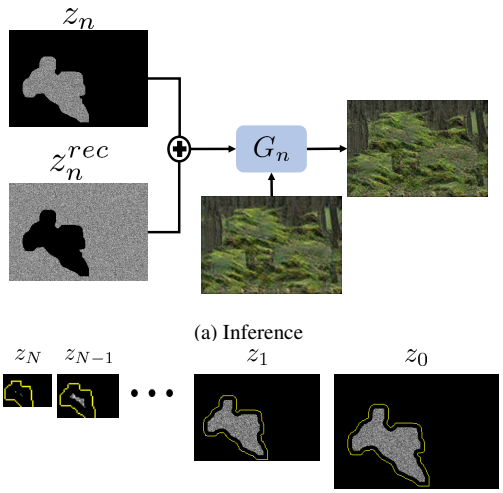


Figure 4. **Patches available for training in each scale.** We want the discriminator to classify only patches not containing missing pixels and thus dilate the mask (yellow) by half a receptive field. Pixels outside the black region are the centers of valid patches. At coarse scales, the number of valid patches is insufficient for training. Therefore, at the coarse scales, we perform naive inpainting rather than masking.



(a) Inference

(b) Sampled area at each scale

Figure 5. **Inference.** (a) At inference time, we sample z_n inside the mask and use z_n^{rec} outside the mask. The combined noise map is injected to the generator together with the up-sampled fake image from the previous scale. (b) The noise z_n is sampled only within regions that do not affect the valid part of the image. This requires eroding the mask (yellow) by half a receptive field.

(not necessarily covering a whole object). As we show, our approach performs at least comparable to the baselines, especially in the task of object removal. This is despite the fact that it has access only the masked image.

4.1. Object removal

The common practice in evaluating generic inpainting methods is to randomly mask regions within images (say from Places [54] or ImageNet [9]) and use the non-masked images as ground-truth. However, this does not simulate well the object removal task, as it results in cutting out of parts of semantic objects. To simulate the object removal task, here we use the Part-Imagenet [12] dataset, which contains images of single objects, along with their corresponding segmentation maps. We use the segmentation map as our mask.

Qualitative comparisons Figure 6 shows completion results generated by our method as well as other inpainting techniques: the non-diverse external methods CA [47] and LAMA [36], the diverse external methods DSI [29], PIC [53], ICT [40] and CoMod-GAN [52], and the non-diverse internal methods Shift-Map [30], Patch-Match [2] and DIP [39]. We use ImageNet models where available (PIC, DSI, and ICT). See SM for comparisons with other models and methods. As can be seen, externally trained methods sometimes generate an object in the missing region, which is undesired in an object-removal setting. Patch-Match and DIP often generate blurry completions. Shift-map fills the missing pixels by copying information from other parts of the image and therefore leads to more realistic results, however it does not offer diverse solutions. Our method generates plausible completions, and suggests diverse solutions (Fig. 7). Note that our method is trained on very limited information (only the non-masked pixels within the image). Yet, it manages to produce diverse completions that are at least comparable in visual quality to diverse external methods.

Human perceptual study We quantify the realism of our results using a user study performed through the Amazon Mechanical Turk platform. We randomly chose 50 images from the Part-Imagenet dataset and conducted 13 surveys, each comparing our method with a different baseline. All images participating in the studies appear in the SM. In each study, we included 50 questions: 5 tutorial trials (with a clear correct answer and feedback to the users) and 45 test questions. Each question displayed the original image, the masked image, and two possible completions: one of our method and one of the competing method. Users were asked to choose which completion they prefer. In total, 50 workers participated in each study. The results are shown in Fig. 8. For the DIP algorithm, we used both the original implementation, which runs for a fixed number of iterations, and an optimized variant that selects the result with the lowest loss along the algorithm's iterations. As can be seen, users tend to prefer our completions over other internal methods (ranking IDC as better than DIP and Patch-Match and comparable to Shift-Map). IDC is also ranked at least on par with most external methods, except for the recent LAMA and CoMod-GAN methods. This is while our method has the advantage of being applicable also to other domains.

Quantitative evaluation We complement the user study with further quantitative evaluation. Since no ground truth is available in the object removal task (*i.e.* we do not have images of the background behid the object), measures like PSNR and FID, which require reference images, cannot be used. We therefore use two common no-reference metrics: the naturalness image quality evaluator (NIQE) [26] and the neural image assessment (NIMA) measure [37]. Results are presented in Tab. 1. As can be seen, in terms of NIQE

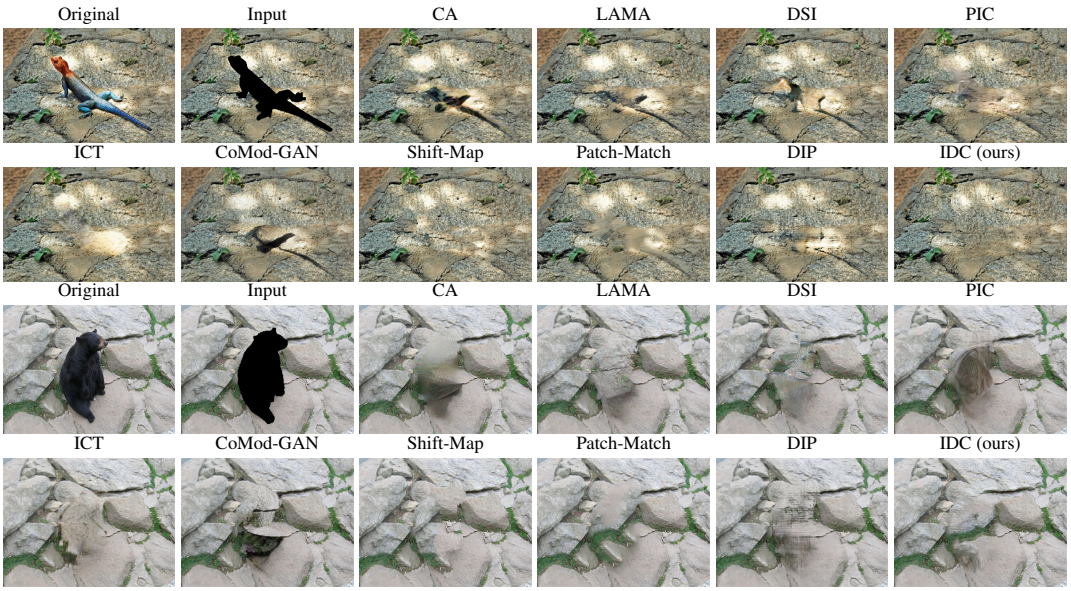


Figure 6. **Qualitative comparison.** Our method is at least comparable to baselines in terms of visual quality, while being the only diverse method that is applicable to arbitrary domains. See SM for completions with other models.

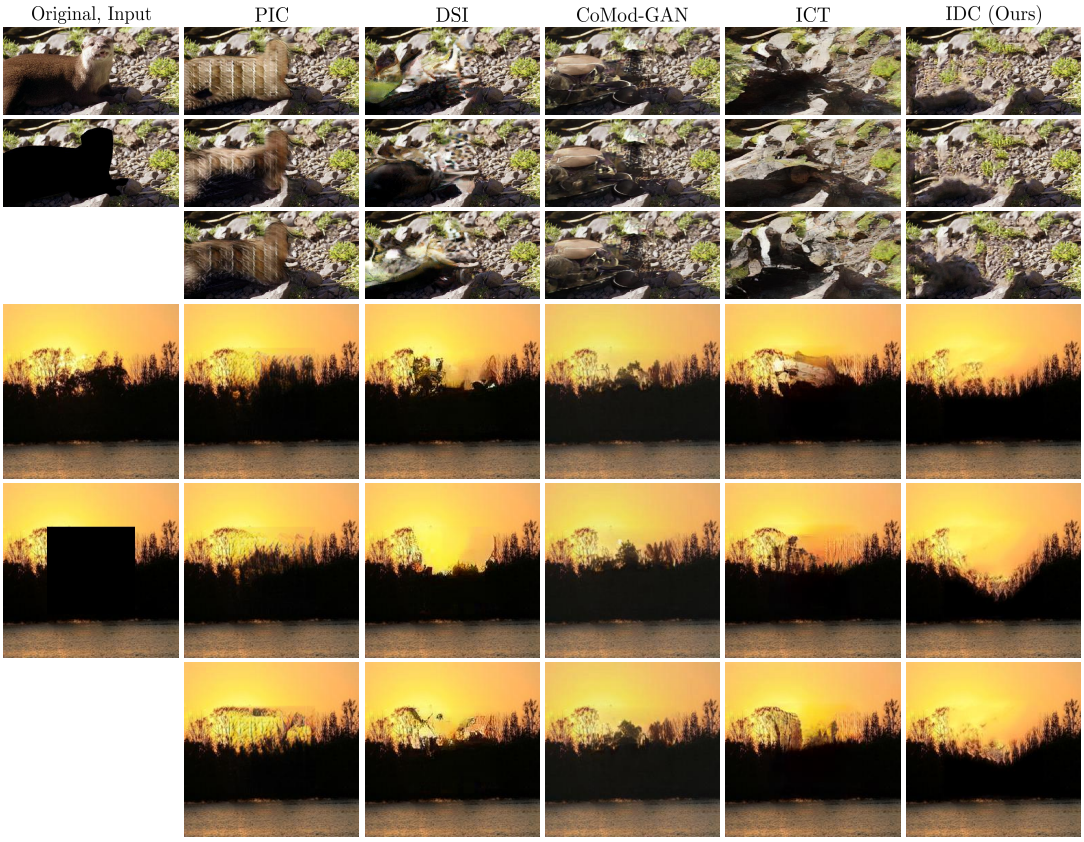


Figure 7. **Internal vs. external diverse completion.** External methods complete the missing region according to the distribution of their training set (we use ImageNet models in the 1st example when available, and Places models in the 2nd). This may result in undesirable effects in the task of object removal.

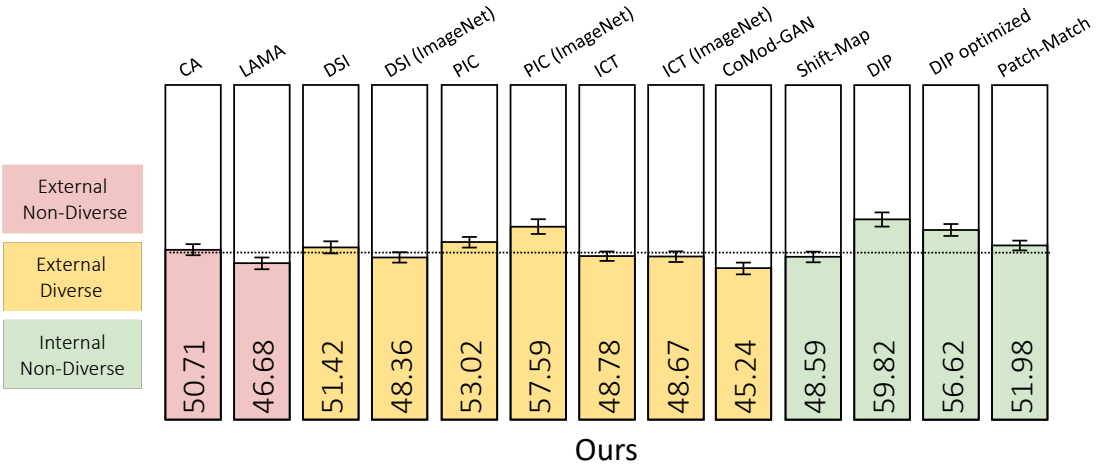


Figure 8. **Human perceptual studies.** We report user preferences for our completions over competing baselines on Part-Imagenet (bootstrap is used for calculating standard deviation). We compare to external models trained on Places dataset and, when available, also to models trained on ImageNet. As can be seen, our method is ranked by users as comparable or better than most baselines, while it is the only diverse technique that is applicable to arbitrary domains (being internally trained).

			NIQE↓	NIMA↑	LPIPS $\cdot 10^{-3}$ (Diversity)
External	Non-Diverse	CA (Places)	5.19	4.81	0
		LAMA (Places)	5.18	4.77	0
	Diverse	DSI (Places/ImageNet)	5.17 / 5.24	4.70 / 4.83	48 / 46
		PIC (Places/ImageNet)	5.15 / 4.82	4.74 / 4.76	23 / 22
		ICT (Places/ImageNet)	5.10 / 5.19	4.78 / 4.77	55 / 62
		CoMod-GAN (Places)	4.91	4.73	13
Internal	Non-Diverse	Shift-Map	5.53	4.81	0
		DIP (original/optimized)	5.54 / 5.25	4.33 / 4.44	0 / 0
		Patch-Match	5.32	4.82	0
	Diverse	IDC (Ours)	5.27	4.79	15

Table 1. **Quantitative evaluation.** We quantify visual quality and semantic diversity on the Part-Imagenet dataset. In external methods, we refer to Places models, and when available, also to ImageNet models. In terms of NIQE (lower is better), our method is ranked as better than all internal methods and slightly inferior to external methods (though this does not always align with users' preferences; see Fig.8). In terms of NIMA (higher is better), our method is ranked as better than DIP, PIC, CoMod-GAN and DSI (Places model) and comparable to all other methods. The semantic diversity achieved by our method is similar to that of CoMod-GAN and somewhat lower than other external diverse inpainting approaches.

(lower is better) our method is better than other internal methods and slightly inferior to external methods, though this does not always align with users' preferences as can be seen in Fig 8. In terms of NIMA (higher is better) our method is ranked higher than DIP, PIC, CoMod-GAN, and DSI (Places model) and on par with all other baselines.

Diversity We also quantify the semantic diversity among the different completions of each method, using the learned perceptual image patch similarity (LPIPS) measure [50]. For each input image, we generate 20 pairs of diverse completions, and calculate the average LPIPS score between all pairs over all images. The results are reported in Tab. 1. As can be seen, our semantic diversity is a bit lower than most external methods. That is, the different completions we generate are not interpreted by a classification network as having very different semantic meanings. However, note that higher semantic diversity is often indicative of insertion of new objects into the missing region, which is not desired in an object completion task (see *e.g.* Fig. 6 and Fig. 7). In any case, it is possible to control the diversity in IDC either by adjusting the noise levels (which has a small effect), or by adjusting the kernel used to erode the mask in each scale. We elaborate on these mechanisms in the SM.

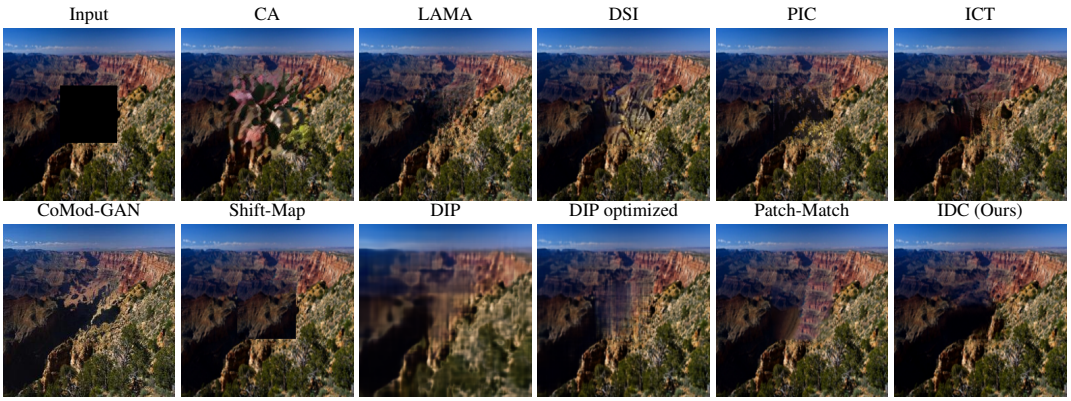


Figure 9. **Qualitative comparison for arbitrary mask completion.** We compare completions of a centered square of size 85×85 pixels. Shift-Map generates a completion with apparent boundaries, and DIP (both variants) and Patch-Match produce blurry outputs. In this specific scenery photo, our method performs at least on par with the external methods.

	Missing part		85×85 centralized square				128×128 centralized square			
			FID↓	NIQE↓	NIMA↑	LPIPS $\cdot 10^{-3}$ (Diversity)	FID↓	NIQE↓	NIMA↑	LPIPS $\cdot 10^{-3}$ (Diversity)
External	Non-Diverse	CA	42.36	4.35	4.63	0	89.17	4.39	4.54	0
		LAMA	29.05	4.58	4.71	0	63.74	4.58	4.69	0
	Diverse	DSI	38.01	4.57	4.58	40	86.05	4.57	4.50	110
		PIC	42.86	4.35	4.69	25	84.64	4.11	4.73	86
		ICT	36.37	4.25	4.71	38	76.85	3.94	4.70	98
		CoMod-GAN	69.44	4.13	4.65	15	102.40	4.06	4.65	39
Internal	Non-Diverse	Shift-Map	45.26	4.34	4.66	0	100.9	4.49	4.54	0
		DIP optimized	85.97	5.80	4.30	0	129.52	5.85	4.05	0
		Patch-Match	44.12	4.58	4.66	0	87.75	4.79	4.63	0
	Diverse	IDC (Ours)	49.9	4.56	4.56	8	97.85	4.57	4.47	60

Table 2. **Quantitative evaluation.** We quantify visual quality and semantic diversity on the Places validation dataset. Our results are slightly inferior to external methods, which were trained specifically on this dataset, but are better than the optimized DIP (which outperforms the original DIP in all scores) and comparable to Shift-Map and Patch-Match.

4.2. Generic inpainting

Although our main goal is object removal, we now provide comparisons to baselines on the task of generic inpainting (*i.e.* inpainting with arbitrary masks). In this setting, the mask may occlude parts of semantic objects, which internal methods cannot complete (see SM). However, as we now show, the performance of our method is usually rather close to externally trained models, even on such tasks. Here we use the Places validation dataset [54], with a missing region of a centralized square of size 85×85 or 128×128. In this setting, since we have ground truth images, we calculate The Fréchet Inception Distance (FID) [13] in addition to NIQE and NIMA. Results are presented in Tab. 2. As can be seen, for both mask sizes, our method achieves a better FID score than DIP and is comparable to Shift-Map and Patch-Match. Note, however, that Shift-map often generates recognizable transitions at boundaries of the inpainted region, which FID does not capture and that the completions of Patch-Match are often blurrier (see Fig. 9). Our results are only slightly inferior to external methods, despite the fact that they were trained on thousands of images from this specific dataset.

5. Conclusions

We presented a method for diverse image completion, which does not require a training set. Since our model is trained only on the non-masked regions within the particular image the user wishes to edit, it is not constrained to any specific image domain or mask location or shape. Our approach can present to the user a variety of different completions with controllable diversity levels. Extensive experiments show that our method is at least comparable to (and sometimes even better than) externally trained models in the task of object removal. Furthermore, for completion of arbitrary regions, which may include parts of semantic objects, our method performs nearly as good as external methods.

References

[1] Coloma Ballester, Marcelo Bertalmio, Vicent Caselles, Guillermo Sapiro, and Joan Verdera. Filling-in by joint interpolation of vector fields and gray levels. *IEEE transactions on image processing*, 10(8):1200–1211, 2001. 1, 2

[2] Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B Goldman. Patchmatch: A randomized correspondence algorithm for structural image editing. *ACM Trans. Graph.*, 28(3):24, 2009. 1, 2, 5

[3] Marcelo Bertalmio, Andrea L Bertozzi, and Guillermo Sapiro. Navier-stokes, fluid dynamics, and image and video inpainting. In *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001*, volume 1, pages I–I. IEEE, 2001. 2

[4] Marcelo Bertalmio, Guillermo Sapiro, Vincent Caselles, and Coloma Ballester. Image inpainting. In *Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 417–424, 2000. 1

[5] Marcelo Bertalmio, Luminita Vese, Guillermo Sapiro, and Stanley Osher. Simultaneous structure and texture image inpainting. *IEEE transactions on image processing*, 12(8):882–889, 2003. 1, 2

[6] Weiwei Cai and Zhanguo Wei. Diversity-generated image inpainting with style extraction. *arXiv preprint arXiv:1912.01834*, 2019. 1

[7] Weiwei Cai and Zhanguo Wei. Piigan: generative adversarial networks for pluralistic image inpainting. *IEEE Access*, 8:48451–48463, 2020. 2

[8] Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on image processing*, 13(9):1200–1212, 2004. 2

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 1, 5

[10] Iddo Drori, Daniel Cohen-Or, and Hezy Yeshurun. Fragment-based image completion. In *ACM SIGGRAPH 2003 Papers*, pages 303–312. 2003. 2

[11] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017. 3

[12] Ju He, Shuo Yang, Shaokang Yang, Adam Kortylewski, Xiaoding Yuan, Jie-Neng Chen, Shuai Liu, Cheng Yang, and Alan Yuille. PartImageNet: A large, high-quality dataset of parts. *arXiv preprint arXiv:2112.00933*, 2021. 5

[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. *Advances in neural information processing systems*, 30, 2017. 8

[14] 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

[15] 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

[16] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 1

[17] Anat Levin, Assaf Zomet, and Yair Weiss. Learning how to inpaint from global image statistics. In *ICCV*, volume 1, pages 305–312, 2003. 2

[18] 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. 3

[19] Yijun Li, Sifei Liu, Jimei Yang, and Ming-Hsuan Yang. Generative face completion. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3911–3919, 2017. 2

[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

[21] Hongyu Liu, Bin Jiang, Yibing Song, Wei Huang, and Chao Yang. Rethinking image inpainting via a mutual encoder-decoder with feature equalizations. In *European Conference on Computer Vision*, pages 725–741. Springer, 2020. 2

[22] Hongyu Liu, Bin Jiang, Yi Xiao, and Chao Yang. Coherent semantic attention for image inpainting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4170–4179, 2019. 2

[23] Hongyu Liu, Ziyu Wan, Wei Huang, Yibing Song, Xintong Han, and Jing Liao. Pd-gan: Probabilistic diverse gan for image inpainting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9371–9381, 2021. 2

[24] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 1

[25] Tomer Michaeli and Michal Irani. Blind deblurring using internal patch recurrence. In *European conference on computer vision*, pages 783–798. Springer, 2014. 4

[26] Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 5

[27] Kamyar Nazeri, Eric Ng, Tony Joseph, Faisal Z Qureshi, and Mehran Ebrahimi. Edgeconnect: Generative image inpainting with adversarial edge learning. *arXiv preprint arXiv:1901.00212*, 2019. 2

[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. 1, 2

[29] Jialun Peng, Dong Liu, Songcen Xu, and Houqiang Li. Generating diverse structure for image inpainting with hierarchical vq-vae. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10775–10784, 2021. 1, 2, 5

[30] Yael Pritch, Eitam Kav-Venaki, and Shmuel Peleg. Shift-map image editing. In *2009 IEEE 12th international conference on computer vision*, pages 151–158. IEEE, 2009. 1, 2, 5

[31] Shuyi Qu, Zhenxing Niu, Kaizhu Huang, Jianke Zhu, Matan Protter, Gadi Zimerman, and Yinghui Xu. Structure first detail next: Image inpainting with pyramid generator. *arXiv preprint arXiv:2106.08905*, 2021. 1, 2

[32] Yurui Ren, Xiaoming Yu, Ruonan Zhang, Thomas H Li, Shan Liu, and Ge Li. Structureflow: Image inpainting via structure-aware appearance flow. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 181–190, 2019. 2

[33] Tamar Rott Shaham, Tali Dekel, and Tomer Michaeli. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4570–4580, 2019. 1, 2

[34] Chitwan Saharia, William Chan, Huiwen Chang, Chris A Lee, Jonathan Ho, Tim Salimans, David J Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. *arXiv preprint arXiv:2111.05826*, 2021. 2

[35] Denis Simakov, Yaron Caspi, Eli Shechtman, and Michal Irani. Summarizing visual data using bidirectional similarity. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2008. 2

[36] Roman Suvorov, Elizaveta Logacheva, Anton Mashikhin, Anastasia Remizova, Arsenii Ashukha, Aleksei Silvestrov, Naejin Kong, Harshith Goka, Kiwoong Park, and Victor Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2149–2159, 2022. 2, 5

[37] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE transactions on image processing*, 27(8):3998–4011, 2018. 5

[38] Alexandru Telea. An image inpainting technique based on the fast marching method. *Journal of graphics tools*, 9(1):23–34, 2004. 1, 2

[39] Dmitry Ulyanov, Andrea Vedaldi, and Victor Lempitsky. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9446–9454, 2018. 2, 4, 5

[40] Ziyu Wan, Jingbo Zhang, Dongdong Chen, and Jing Liao. High-fidelity pluralistic image completion with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4692–4701, 2021. 1, 2, 5

[41] Junyuan Xie, Linli Xu, and Enhong Chen. Image denoising and inpainting with deep neural networks. *Advances in neural information processing systems*, 25, 2012. 1

[42] Shunxin Xu, Dong Liu, and Zhiwei Xiong. E2i: Generative inpainting from edge to image. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(4):1308–1322, 2020. 2

[43] Zongben Xu and Jian Sun. Image inpainting by patch propagation using patch sparsity. *IEEE transactions on image processing*, 19(5):1153–1165, 2010. 1, 2

[44] Chao Yang, Xin Lu, Zhe Lin, Eli Shechtman, Oliver Wang, and Hao Li. High-resolution image inpainting using multi-scale neural patch synthesis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6721–6729, 2017. 2

[45] Raymond Yeh, Chen Chen, Teck Yian Lim, Mark Hasegawa-Johnson, and Minh N Do. Semantic image inpainting with perceptual and contextual losses. *arXiv preprint arXiv:1607.07539*, 2(3), 2016. 2

[46] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 1

[47] Jiahui Yu, Zhe Lin, Jimei Yang, Xiaohui Shen, Xin Lu, and Thomas S Huang. Generative image inpainting with contextual attention. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5505–5514, 2018. 1, 2, 5

[48] 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/CVF International Conference on Computer Vision*, pages 4471–4480, 2019. 2

[49] Yingchen Yu, Fangneng Zhan, Rongliang Wu, Jianxiong Pan, Kaiwen Cui, Shijian Lu, Feiying Ma, Xuansong Xie, and Chunyan Miao. Diverse image inpainting with bidirectional and autoregressive transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 69–78, 2021. 2

[50] 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

[51] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5741–5750, 2020. 1, 2

[52] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021. 1, 2, 5

[53] 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, 5

[54] Bolei Zhou, Agata Lapedriza, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. Learning deep features for scene recognition using places database. *Advances in neural information processing systems*, 27, 2014. 1, 5, 8