International Journal of Computer Vision (2020) 128:2363–2365
https://doi.org/10.1007/s11263-020-01380-5

Guest Editorial: Generative Adversarial Networks for Computer Vision

Jun-Yan Zhu[1] · Hongsheng Li[2] · Eli Shechtman[1] · Ming-Yu Liu[3] · Jan Kautz[3] · Antonio Torralba[4]

Received: 3 September 2020 / Published online: 14 September 2020

For decades, the field of computer vision has used generative models for both recognition and synthesis tasks. For example, seminal work adopted probabilistic generative models for image classification (Weber et al. 2000; Fergus et al. 2003; Fei-Fei and Perona 2005), shape perception (Freeman 1994) and digit recognition (Revow et al. 1996; Learned-Miller 2005). Meantime, classic generative models, such as Gaussian Mixture Model and principal components, have long been used to learn prior models for image restoration (Olshausen and Field 1996; Portilla and Simoncelli 2000; Zoran and Weiss 2011), segmentation (Rother et al. 2004), and face modeling (Blanz and Vetter 1999; Cootes et al. 2001). Unfortunately, due to the limited capacity, these models either learn local image statistics of pixel values, gradients, and feature descriptors, or only work well on aligned objects such as digits and faces. None of the above models are able to learn the distribution of in-the-wild natural images and capture long-range dependence beyond local regions. As a result, the above work mostly focused on low-level vision and graphics applications. For recognition tasks, classic generative models seldom outperformed discriminative classifiers. For synthesis tasks, these methods struggled to synthesize natural images with the same expressiveness and fidelity as 3D graphics rendering pipelines, with the notable exception of photorealistic face synthesis with Morphable Models (Blanz and Vetter 1999) and Active Appearance Models (Cootes et al. 2001).

Recently, a wide range of deep generative models (Hinton and Salakhutdinov 2006; Goodfellow et al. 2014; Kingma and Welling 2014; Dinh et al. 2016; Van den Oord et al. 2016) have been developed for modeling the distribution of full images. Among them, Generative Adversarial Networks (GANs) (Goodfellow et al. 2014) have been at the forefront of research in the past few years, producing high-quality images while enabling efficient inference. GANs can approximate real data distributions and synthesize realistic data samples. The learning algorithm is carried through a two-player game between a generator that synthesizes an image, and a discriminator that distinguishes real images from synthetic ones. Compared to prior work (Tu 2007), the success of GANs partly comes from high-capacity CNN classifiers (Krizhevsky e al. 2012) that can easily detect fake samples and expressive upsampling networks (Dosovitskiy et al. 2015) that can output high-dimensional images from low-dimensional vectors. For the first time ever, computer vision and graphics community is given a generative model capable of modeling the complexity and realism of natural images by learning hierarchical feature representations from high-level object concepts to low-level visual cues (Bau et al. 2019). Moreover, for certain object categories, recent GANs (Karras et al. 2020) can achieve similar or sometimes better image quality, compared to traditional 3D rendering pipelines.

What could we do with this powerful new computational tool? It turns out that its application is not limited to generating samples from certain data distributions but also has inspired many other research trends, including image generation and editing, feature learning, visual domain adaptation, data generation, and augmentation for visual recognition, often leading to state-of-the-art results. While GANs have achieved substantial progress for various computer vision applications, many issues remain to be solved, and new

✉ Jun-Yan Zhu
junyanz@cs.cmu.edu

Hongsheng Li
hsli@ee.cuhk.edu.hk

Eli Shechtman
elishe@adobe.com

Ming-Yu Liu
mingyul@nvidia.com

Jan Kautz
jkautz@nvidia.com

Antonio Torralba
torralba@csail.mit.edu

[1] Adobe Research, San Jose, CA, USA

[2] The Chinese University of Hong Kong, Hong Kong SAR, China

[3] NVIDIA Research, Santa Clara, CA, USA

[4] MIT CSAIL, Cambridge, MA, USA

research problems emerge. For example, what are the effective network structures and objective functions for generating different visual data (e.g., images, videos, 3D)? What are the proper metrics for evaluating deep generative models? How can we improve the photorealism and resolution of the synthesized data samples? How can the generated data help solve downstream computer vision tasks?

The goal of this special issue is to solicit original work at the intersection of computer vision and deep generative models such as GANs. This special issue received 58 initial submissions. Five submissions were either withdrawn or were rejected after abstract review, 53 manuscripts went through the IJCV review cycle, out of which 21 papers were accepted to this special issue. These papers spanned the following topics:

- *Image-to-Image Translation:* Several works explore new directions in Conditional GANs, including compositionality, disentanglement, and robustness. Compositional GAN (s11263-020-01336-9) presents a generative model that can composite a pair of objects with consistent relative scaling, spatial layout, occlusion, and viewpoint. DRIT++ (s11263-019-01284-z) propose a cross-cycle consistency loss to enable the disentanglement of content and domain during image-to-image translation. RoC-GAN (s11263-020-01348-5) improves the robustness of the conditional GANs by encouraging the generator's outputs to stay on the image manifold of target domain. Layout2Image (s11263-020-01300-7) learns to synthesize realistic images given object labels and bounding boxes, by disentangling an object into the semantic category and appearance representation. DRPAN (s11263-019-01273-2) enables high-quality image-to-image translation by *revising* unrealistic regions, given the feedback from the discriminator. Finally, a new type of conditional generative model, Implicit Maximum Likelihood Estimation (s11263-020-01325-y), is proposed and compared against several conditional GANs baselines.
- *Video Synthesis:* This special issue also include two image-to-video generation methods. pix2vid (s11263-020-01334-x) synthesizes a short video given a single structure annotation, while Zhao et al. (s11263-020-01328-9) propose generating an output video given a reference video's motion and input images' appearance. Several applications have been explored, including facial expression retargeting, human pose forecasting, and video prediction.
- *3D-Aware GANs:* A few papers go beyond 2D image synthesis and incorporate 3D structure into the generative models. Ververas and Zafeiriou (s11263-020-01338-7) propose an image-to-image translation model that transforms face images conditioned on continuous 3D blendshape models. Gadelha et al. (s11263-020-01335-w) learn a GAN model of 3D shapes in a voxel representation, purely from 2D image observations. Pix2Shape (s11263-020-01322-1) tackles the same problem setting as above, but with a different 3D representation – a view-dependent explicit surfel representation. This allows the model to efficiently sample scene information. To create facial image manipulation effects, Geng et al. (s11263-020-01361-8) disentangle a face image into texture, shape, and identity, using a 3D face fitting model, while 3DFaceGAN (s11263-020-01329-8) learns a GAN model of 3D facial shapes, with applications on 3D face translation and synthesis.
- *Visual Recognition with GANs:* Several works propose improving recognition systems through adversarial data augmentation and domain adaptation. For examples, Dutta and Akata (s11263-020-01350-x) use semantically aligned paired cycle-consistent adversarial networks for any-shot image retrieval. Nie and Shen (s11263-020-01321-2) propose using confidence information provided by the adversarial network to enhance the design of a supervised segmentation network. To improve fine-grained recognition systems, where annotated data is scarce, Yu and Grauman (s11263-020-01344-9) use attribute-conditional generative models to *densify* the space of training images. Wu et al. (s11263-020-01291-5) use adversarial learning to improve the robustness of handwritten mathematical expression systems with respect to different writing styles.
- *Inverting GANs:* To edit a real image using unconditional GANs, one needs to first project the image into the latent space of GANs (Zhu et al. 2016). Despite recent efforts, it still remains challenging and computationally-expensive for deep generators and images in the wild. To tackle these issues, Band et al. (s11263-020-01311-4) propose to reuse the discriminator's feature representation as part of the encoder. This improves projection with minimal training overhead. MimicGAN (s11263-020-01310-5) proposes modeling common image corruptions, such as cropping rotation, missing pixels, during the projection, expanding the scope of images that can be possibly embedded.
- *Theory and Training Method:* Abbasnejad et al. (s11263-020-01360-9) propose a Generative Adversarial Density Estimator aiming to bridge the gap between maximum likelihood approaches and likelihood-free approaches on density estimation. Saito et al. (s11263-020-01333-y) present a memory-efficient method for unsupervised learning of high-resolution video generation. The computational cost scales only linearly with the resolution.

Finally, we would like to thank all the reviewers for their helpful and constructive comments. We are excited by the outcome and hope that readers enjoy the insights in the articles.

References

Bau, D., Zhu, J. Y., Strobelt, H., Bolei, Z., Tenenbaum, J. B., Freeman, W. T., & Torralba, A. (2019) Gan dissection: Visualizing and understanding generative adversarial networks. In *International conference on learning representations (ICLR)*.

Blanz, V., & Vetter, T. (1999). A morphable model for the synthesis of 3d faces. In *ACM SIGGRAPH*.

Cootes, T. F., Edwards, G. J., & Taylor, C. J. (2001). Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *23*(6), 681–685.

Dinh, L., Sohl-Dickstein, J., & Bengio, S. (2016). Density estimation using real nvp. arXiv preprint. arXiv:1605.08803.

Dosovitskiy, A., Tobias Springenberg, J., & Brox, T. (2015). Learning to generate chairs with convolutional neural networks. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Fei-Fei, L., & Perona, P. (2005). A Bayesian hierarchical model for learning natural scene categories. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Fergus, R., Perona, P., & Zisserman, A. (2003). Object class recognition by unsupervised scale-invariant learning. In *IEEE conference on computer vision and pattern recognition (CVPR)*. vol. 2, pp. II–II.

Freeman, W. T. (1994). The generic viewpoint assumption in a framework for visual perception. *Nature*, *368*(6471), 542–545.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*.

Hinton, G. E., & Salakhutdinov, R. R. (2006). Reducing the dimensionality of data with neural networks. *Science*, *313*(5786), 504–507.

Karras, T., Laine, S., Aittala, M., Hellsten, J., Lehtinen, J., & Aila, T. (2020). Analyzing and improving the image quality of stylegan. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. *International conference on learning representations (ICLR)*.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*.

Learned-Miller, E. G. (2005). Data driven image models through continuous joint alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(2), 236–250.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, *381*, 607–609.

Portilla, J., & Simoncelli, E. P. (2000). A parametric texture model based on joint statistics of complex wavelet coefficients. *International Journal of Computer Vision (IJCV)*, *40*(1), 49–70.

Revow, M., Williams, C. K., & Hinton, G. E. (1996). Using generative models for handwritten digit recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *18*(6), 592–606.

Rother, C., Kolmogorov, V., & Blake, A. (2004). Grabcut interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics (TOG)*, *23*(3), 309–314.

Tu, Z. (2007). Learning generative models via discriminative approaches. In *IEEE conference on computer vision and pattern recognition (CVPR)*.

Van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al. (2016). Conditional image generation with pixelcnn decoders. In *Advances in neural information processing systems*.

Weber, M., Welling, M., & Perona, P. (2000). Towards automatic discovery of object categories. *IEEE conference on computer vision and pattern recognition (CVPR)*, vol. 2, pp. 101–108.

Zhu, J. Y., Krähenbühl, P., Shechtman, E., & Efros, A. A. (2016). Generative visual manipulation on the natural image manifold. In *European conference on computer vision (ECCV)*.

Zoran, D., & Weiss, Y. (2011). From learning models of natural image patches to whole image restoration. In *IEEE international conference on computer vision (ICCV)*.