Instance-Conditioned GAN

Arantxa Casanova et al.

Facebook AI Research

NeurIPS 2021

Presented by Minho Park

Modeling Complex Distributions

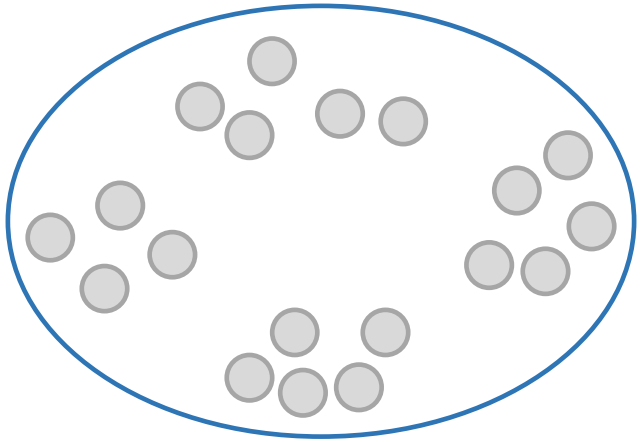
- Modeling complex distributions remains challenging in unconditional settings.
  - E.g., ImageNet and COCO-Stuff
- Unconditional GANs.
- Class conditional GANs.
  - Provide higher quality samples than unconditional counterparts.
  - Require labeled data which may be unavailable or costly to obtain.

Modeling Complex Distributions

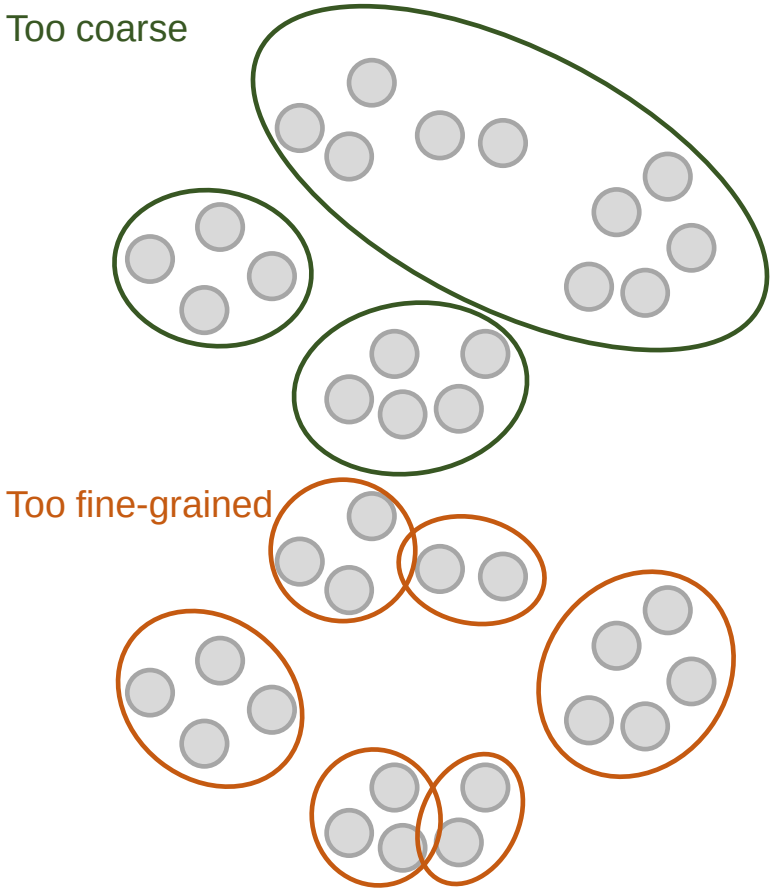
- Unsupervised data partitioning.
- Coarse and non-overlapping data partitions.
  - Partitions often contain different types of objects or scenes.
  - The diversity of data points may result in a manifold with low density regions.
- Finer partitions.
  - The clusters may contain too few data points.
  - The generator and discriminator needs more data points to properly model their data distribution.

Modeling Complex Distributions

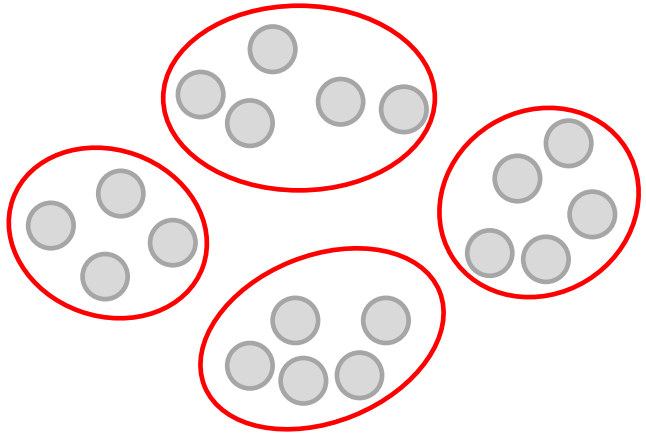
- Comparison.



Unconditional GANs



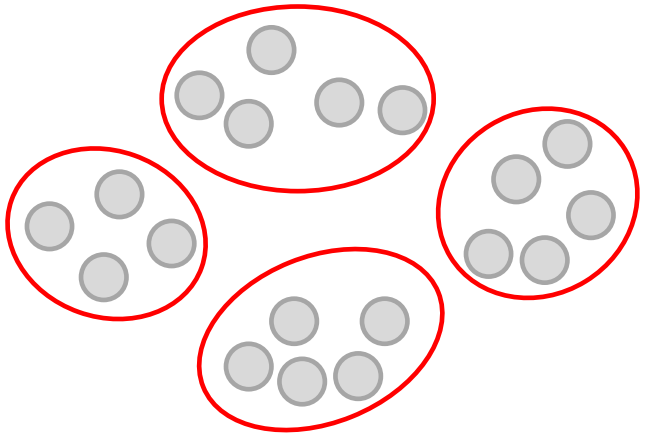
Un-/Self-supervised partitions



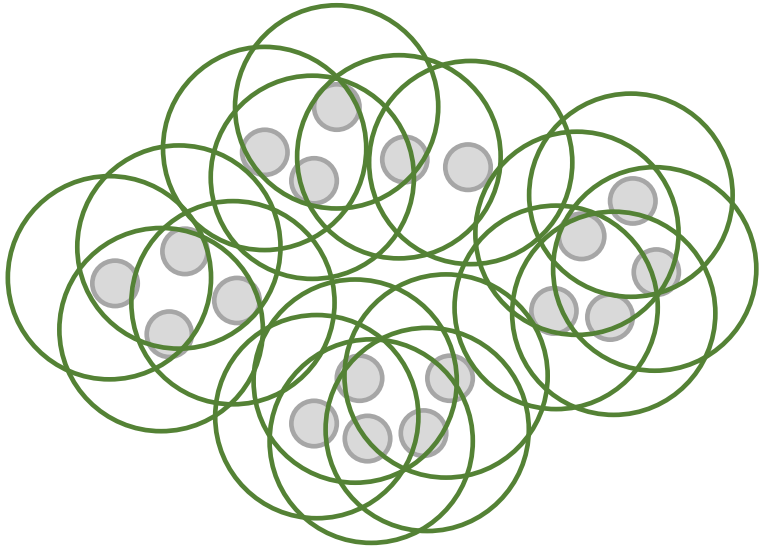
Class conditional GANs

Modeling Complex Distributions

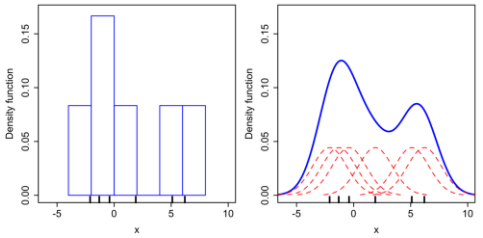
- IC-GAN can be seen as a mixture density estimator (e.g., kernel density estimation).
- Unlike KDE, IC-GAN does not model the data likelihood explicitly.



Class conditional GANs



Instance-conditioned GAN



Comparison of the histogram (left) and kernel density estimate (right) constructed using the same data.

Instance-conditioned GAN

- Objective: Model the underlying data distribution $p(x)$

  as a mixture of conditional distribution $p(x|h_i)$

  around each of M instance feature vectors h_i in the dataset, such that

$$p(x) \approx \frac{1}{M}\sum_i p(x|h_i)$$

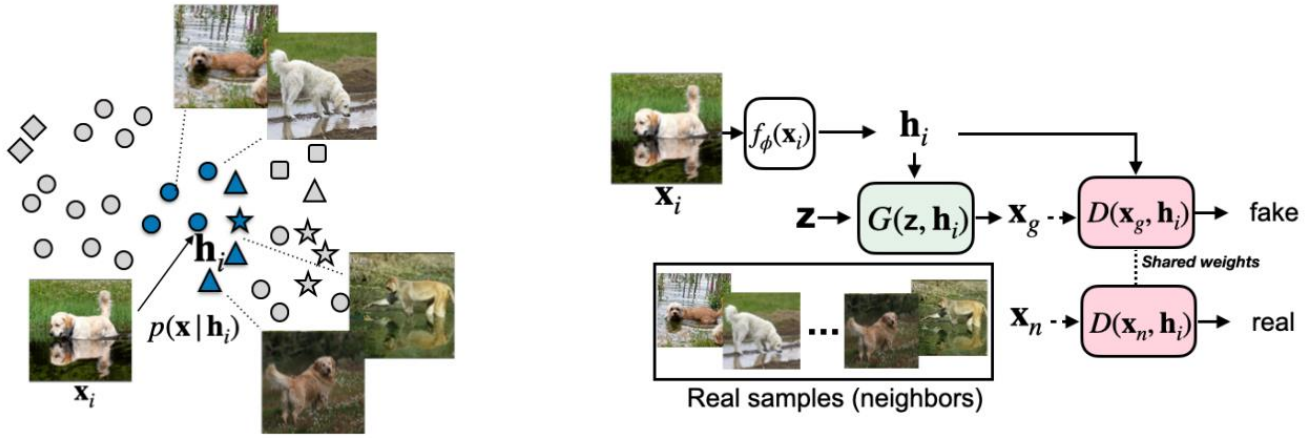
  where h_i is extracted by un-/self-supervised encoder $f_\phi(\cdot)$, i.e., $f_\phi(x_i) = h_i$.

- However, instead of modeling the data likelihood explicitly, IC-GAN implicitly modeling with a GAN.

Instance-conditioned GAN

- Model $p(x|h_i)$ with a GAN, i.e., $G_{\theta_G}(z, h_i), D_{\theta_D}(x, h_i)$.

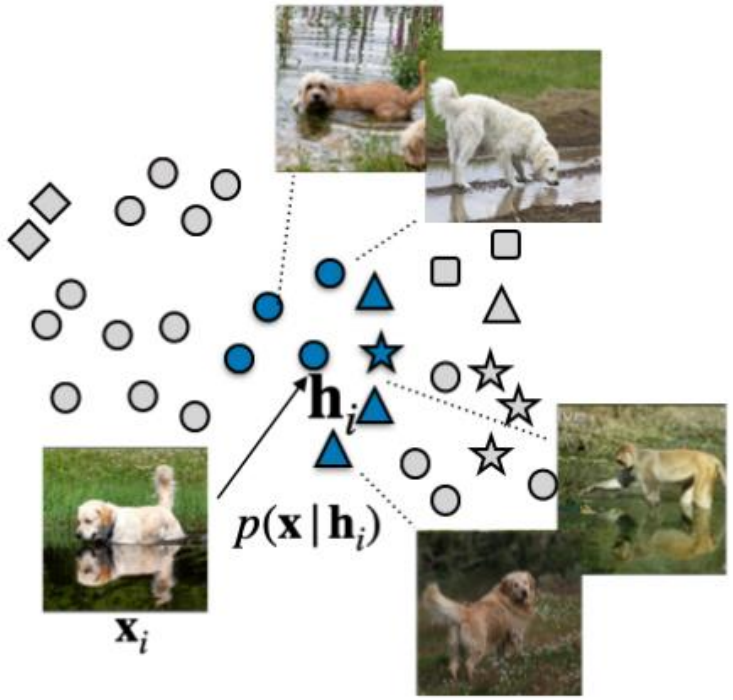
$$\min_G \max_D \mathbb{E}_{\mathbf{x}_i \sim p(\mathbf{x}), \mathbf{x}_n \sim \mathcal{U}(\mathcal{A}_i)} [\log D(\mathbf{x}_n, f_\phi(\mathbf{x}_i))] + \mathbb{E}_{\mathbf{x}_i \sim p(\mathbf{x}), \mathbf{z} \sim p(\mathbf{z})} [\log(1 - D(G(\mathbf{z}, f_\phi(\mathbf{x}_i)), f_\phi(\mathbf{x}_i)))].$$



(a) Neighborhood $\mathcal{A}_i$ of instance $\mathbf{h}_i$ (b) Schematic illustration of the IC-GAN workflow

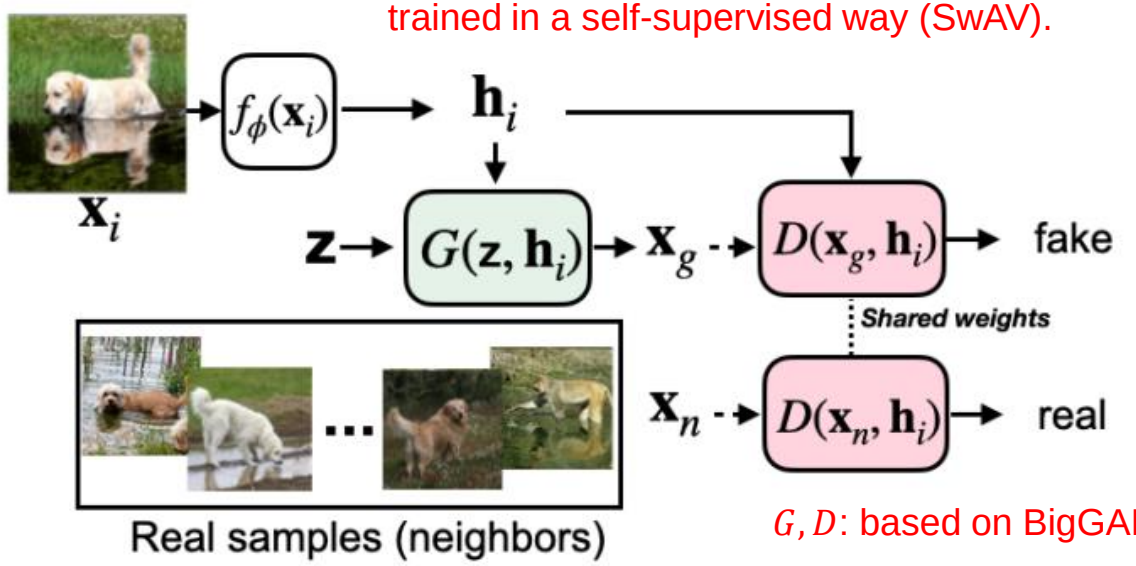
Figure 2: Overview of IC-GAN. (a) The goal of the generator is to generate realistic images similar to the neighbors of $\mathbf{h}_i$, defined in the embedding space using cosine similarity. Five out of seven neighbors are shown in the figure. Note that images in the same neighborhood may belong to different classes (depicted as different shapes). (b) Conditioned on instance features $\mathbf{h}_i$ and noise $\mathbf{z}$, the generator produces a synthetic sample $\mathbf{x}_g$. Generated samples and real samples (neighbors of $\mathbf{h}_i$) are fed to the discriminator, which is conditioned on the same $\mathbf{h}_i$.

Instance-conditioned GAN



(a) Neighborhood $\mathcal{A}_i$ of instance $\mathbf{h}_i$

$f_\phi(\cdot)$: pretrained ResNet50
trained in a self-supervised way (SwAV).



G, D: based on BigGAN

(b) Schematic illustration of the IC-GAN workflow

Distance metric: cosine similarity
Number of neighbors is a hyperparameter (50 for IN, 5 for COCO-Stuff).

Instance-conditioned GAN

- Qualitative results on unlabeled IN.

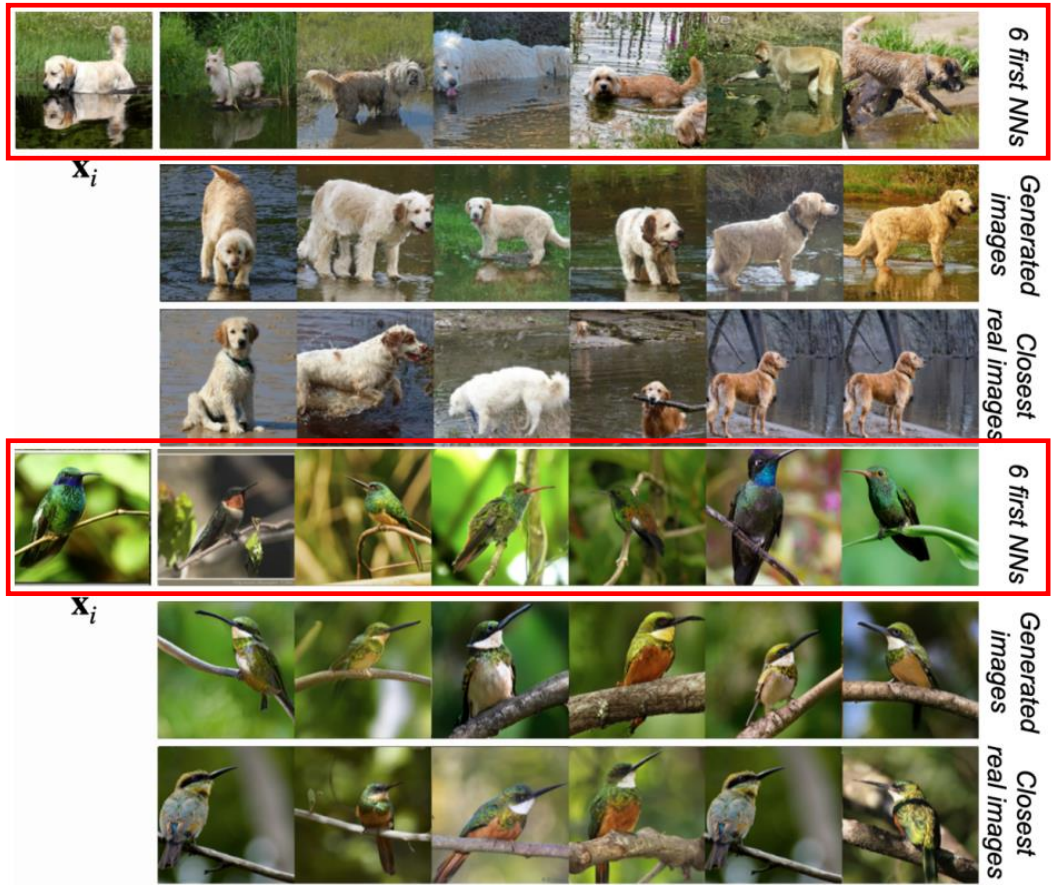


Figure 5: Qualitative results on unlabeled ImageNet (256×256). Next to each input sample $\mathbf{x}_i$, used to obtain the instance features $\mathbf{h}_i = f_\theta(\mathbf{x}_i)$, the six nearest neighbors in the feature space of f_θ are displayed. Below the neighbors, generated images sampled from IC-GAN with a BigGAN backbone and conditioned on $\mathbf{h}_i$ are depicted. Immediately below the generated images, the closest image in the ImageNet training set is shown for each example (cosine distance in the feature space of f_θ).

Extension to Class-conditional Generation

- Conditioning the generator and discriminator on a class label y.
- Train the feature extractor $f_\phi(\cdot)$ for the classification task on ImageNet.
- **Model $p(x|h_i, y_j)$ with a generator $G_{\theta_G}(z, h_i, y_j)$ trained jointly with a discriminator $D_{\theta_D}(x, h_i, y_j)$.**

(a) IC-GAN samples (b) Class-conditional IC-GAN samples

Experiments in Unlabeled Setting

Table 1: Results for ImageNet in unlabeled setting. For fair comparison with [42] at 64×64 resolution, we trained an unconditional BigGAN model and report the non-official FID and IS scores – computed with Pytorch rather than TensorFlow – indicated with *. †: increased parameters to match IC-GAN capacity. DA: 50% horizontal flips in (**d**) real and fake samples, and (**i**) conditioning instances. $ch\times$: Channel multiplier that affects network width as in BigGAN.

Method	**Res.**	**↓FID**	**↑IS**
Self-sup. GAN [42]	64	19.2*	16.5*
Uncond. BigGAN†	64	16.9* ± 0.0	14.6* ± 0.1
IC-GAN	64	10.4* ± 0.1	21.9* ± 0.1
IC-GAN + DA (**d**,**i**)	64	**9.2*** ± 0.0	**23.5*** ± 0.1
MGAN [23]	128	58.9	13.2
PacGAN2 [32]	128	57.5	13.5
Logo-GAN-AE [46]	128	50.9	14.4
Self-cond. GAN [33]	128	41.7	14.9
Uncond. BigGAN [36]	128	25.3	20.4
SS-cluster GAN [36]	128	22.0	23.5
PGMGAN [2]	128	21.7	23.3
IC-GAN	128	13.2 ± 0.0	45.5 ± 0.2
IC-GAN + DA (**d**,**i**)	128	**11.7** ± 0.0	**48.7** ± 0.1
ADM [12]	256	32.5	37.6
IC-GAN ($ch \times 64$)	256	17.0 ± 0.2	53.0 ± 0.4
IC-GAN ($ch \times 64$) + DA (**d**,**i**)	256	17.4 ± 0.1	53.5 ± 0.5
IC-GAN ($ch \times 96$) + DA (**d**)	256	**15.6** ± 0.1	**59.0** ± 0.4

Table 2: Quantitative results on COCO-Stuff. IC-GAN trained on ImageNet indicated as "transf". Some non-zero standard deviations are reported as 0.0 because of rounding.

			↓FID			**↑ LPIPS**
128×128	**# prms.**	**train**	**eval**	**eval seen**	**eval unseen**	**eval**
LostGANv2 [49]	41 M	12.8 ± 0.1	40.7 ± 0.3	80.0 ± 0.4	55.2 ± 0.5	0.45 ± 0.1
OC-GAN [50]	170 M	—	45.1 ± 0.3	85.8 ± 0.5	60.1 ± 0.2	0.13 ± 0.1
Unconditional (BigGAN)	18 M	17.9 ± 0.1	46.9 ± 0.5	103.8 ± 0.8	60.9 ± 0.7	0.68 ± 0.1
IC-GAN (BigGAN)	22 M	16.8 ± 0.1	44.9 ± 0.5	81.5 ± 1.3	60.5 ± 0.5	0.67 ± 0.1
IC-GAN (BigGAN, transf.)	77 M	**8.5** ± 0.0	**35.6** ± 0.2	77.0 ± 1.0	**48.9** ± 0.2	**0.69** ± 0.1
Unconditional (StyleGAN2)	23 M	8.8 ± 0.1	37.8 ± 0.2	92.1 ± 1.0	53.2 ± 0.5	0.68 ± 0.1
IC-GAN (StyleGAN2)	24 M	8.9 ± 0.0	36.2 ± 0.2	**74.3** ± 0.8	50.8 ± 0.3	0.67 ± 0.1
256×256						
LostGANv2 [49]	46 M	18.0 ± 0.1	47.6 ± 0.4	88.5 ± 0.4	62.0 ± 0.6	0.56 ± 0.1
OC-GAN [50]	190 M	—	57.0 ± 0.1	98.7 ± 1.2	71.4 ± 0.5	0.21 ± 0.1
Unconditional (BigGAN)	21 M	51.0 ± 0.1	81.6 ± 0.5	135.1 ± 1.6	95.8 ± 1.1	**0.77** ± 0.1
IC-GAN (BigGAN)	26 M	24.6 ± 0.1	53.1 ± 0.4	88.5 ± 1.8	69.1 ± 0.6	0.73 ± 0.1
IC-GAN (BigGAN, transf.)	90 M	13.9 ± 0.1	**40.9** ± 0.3	79.4 ± 1.2	**55.6** ± 0.6	0.76 ± 0.1
Unconditional (StyleGAN2)	23 M	**7.1** ± 0.0	44.6 ± 0.4	98.1 ± 1.7	59.9 ± 0.5	0.76 ± 0.1
IC-GAN (StyleGAN2)	25 M	9.6 ± 0.0	41.4 ± 0.2	**76.7** ± 0.6	57.5 ± 0.5	0.74 ± 0.1

(a) $\mathbf{x}_i$ (b) IC-GAN (StyleGAN2) (c) IC-GAN (BigGAN) (d) IC-GAN (BigGAN, transf.)

Figure 3: Qualitative comparison for scene generation on 256×256 COCO-Stuff.

Off-the-shelf Transfer to Other Datasets

- IN pretrained IC-GAN outperforms COCO-Stuff pretrained IC-GAN on COCO-Stuff.
- FID between real IN and real COCO-Stuff is 37.2.
- **Remarkable transfer capabilities of IC-GAN!**

Table 2: Quantitative results on COCO-Stuff. IC-GAN trained on ImageNet indicated as "transf". Some non-zero standard deviations are reported as 0.0 because of rounding.

			↓FID			↑ LPIPS
128×128	# prms.	train	eval	eval seen	eval unseen	eval
LostGANv2 [49]	41 M	12.8 ± 0.1	40.7 ± 0.3	80.0 ± 0.4	55.2 ± 0.5	0.45 ± 0.1
OC-GAN [50]	170 M	—	45.1 ± 0.3	85.8 ± 0.5	60.1 ± 0.2	0.13 ± 0.1
Unconditional (BigGAN)	18 M	17.9 ± 0.1	46.9 ± 0.5	103.8 ± 0.8	60.9 ± 0.7	0.68 ± 0.1
IC-GAN (BigGAN)	22 M	16.8 ± 0.1	44.9 ± 0.5	81.5 ± 1.3	60.5 ± 0.5	0.67 ± 0.1
IC-GAN (BigGAN, transf.)	77 M	**8.5** ± 0.0	**35.6** ± 0.2	77.0 ± 1.0	**48.9** ± 0.2	**0.69** ± 0.1
Unconditional (StyleGAN2)	23 M	8.8 ± 0.1	37.8 ± 0.2	92.1 ± 1.0	53.2 ± 0.5	0.68 ± 0.1
IC-GAN (StyleGAN2)	24 M	8.9 ± 0.0	36.2 ± 0.2	**74.3** ± 0.8	50.8 ± 0.3	0.67 ± 0.1
256×256						
LostGANv2 [49]	46 M	18.0 ± 0.1	47.6 ± 0.4	88.5 ± 0.4	62.0 ± 0.6	0.56 ± 0.1
OC-GAN [50]	190 M	—	57.0 ± 0.1	98.7 ± 1.2	71.4 ± 0.5	0.21 ± 0.1
Unconditional (BigGAN)	21 M	51.0 ± 0.1	81.6 ± 0.5	135.1 ± 1.6	95.8 ± 1.1	**0.77** ± 0.1
IC-GAN (BigGAN)	26 M	24.6 ± 0.1	53.1 ± 0.4	88.5 ± 1.8	69.1 ± 0.6	0.73 ± 0.1
IC-GAN (BigGAN, transf.)	90 M	13.9 ± 0.1	**40.9** ± 0.3	79.4 ± 1.2	**55.6** ± 0.6	0.76 ± 0.1
Unconditional (StyleGAN2)	23 M	**7.1** ± 0.0	44.6 ± 0.4	98.1 ± 1.7	59.9 ± 0.5	0.76 ± 0.1
IC-GAN (StyleGAN2)	25 M	9.6 ± 0.0	41.4 ± 0.2	**76.7** ± 0.6	57.5 ± 0.5	0.74 ± 0.1

(c) IC-GAN transfer samples

(d) Class-conditional IC-GAN transfer samples

Experiments in Class-conditional Setting

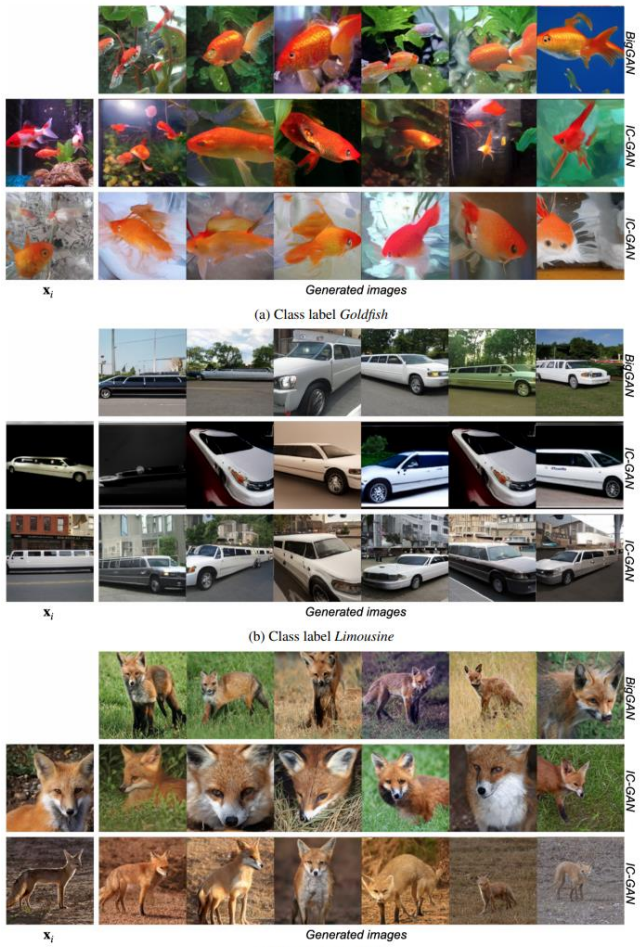
Table 3: Class-conditional results on ImageNet. *: Trained using open source code. DA: 50% horizontal flips in (**d**) real and fake samples, and (**i**) conditioning instances. $ch\times$: Channel multiplier that affects network width. †: numbers from the original paper, as training diverged with the BigGAN opensourced code.

	Res.	↓FID	↑IS
BigGAN* [5]	64	12.3 ± 0.0	27.0 ± 0.2
BigGAN* [5] + DA (**d**)	64	10.2 ± 0.1	30.1 ± 0.1
IC-GAN	64	8.5 ± 0.0	39.7 ± 0.2
IC-GAN + DA(**d**, **i**)	64	**6.7** ± 0.0	**45.9** ± 0.3
BigGAN* [5]	128	9.4 ± 0.0	98.7 ± 1.1
BigGAN* [5] + DA(**d**)	128	**8.0** ± 0.0	107.2 ± 0.9
IC-GAN	128	10.6 ± 0.1	100.1 ± 0.5
IC-GAN + DA(**d**, **i**)	128	9.5 ± 0.1	**108.6** ± 0.7
BigGAN* [5] ($ch \times 64$)	256	8.0 ± 0.1	139.1 ± 0.3
BigGAN* [5] ($ch \times 64$) + DA(**d**)	256	8.3 ± 0.1	125.0 ± 1.1
IC-GAN ($ch \times 64$)	256	8.3 ± 0.1	143.7 ± 1.1
IC-GAN ($ch \times 64$) + DA(**d**, **i**)	256	**7.5** ± 0.0	152.6 ± 1.1
BigGAN† [5] ($ch \times 96$)	256	8.1	144.2
IC-GAN ($ch \times 96$) + DA(**d**)	256	8.2 ± 0.1	**173.8** ± 0.9

Table 4: Class-conditional results on ImageNet-LT. *: Trained using open source code.

	Res.	↓train FID	↑train IS	↓val FID	many/med/few ↓val FID	↑val IS
BigGAN* [5]	64	27.6 ± 0.1	18.1 ± 0.2	28.1 ± 0.1	28.8 / 32.8 / 48.4 ± 0.2	16.0 ± 0.1
IC-GAN	64	**23.2** ± 0.1	**19.5** ± 0.1	**23.4** ± 0.1	**23.8 / 28.0 / 42.7** ± 0.1	**17.6** ± 0.1
BigGAN* [5]	128	31.4 ± 0.1	30.6 ± 0.1	35.4 ± 0.1	34.0 / 43.5 / 64.4 ± 0.2	24.9 ± 0.2
IC-GAN	128	**23.4** ± 0.1	**39.6** ± 0.2	**24.9** ± 0.1	**24.3 / 31.4 / 53.6** ± 0.3	**32.5** ± 0.1
BigGAN* [5]	256	27.8 ± 0.0	58.2 ± 0.2	31.4 ± 0.1	28.1 / 40.9 / 67.6 ± 0.3	44.7 ± 0.2
IC-GAN	256	**21.7** ± 0.1	**66.5** ± 0.3	**23.4** ± 0.1	**20.6 / 32.4 / 60.0** ± 0.2	**51.7** ± 0.1

Additional Qualitative Results



(a) Class label *Goldfish*

(b) Class label *Limousine*

(c) Class label *Red fox*

Figure 8: Qualitative results in 256×256 ImageNet. For each class, generated images with BigGAN are presented in the first row, while the second and third row show generated images using class-conditional IC-GAN with a BigGAN backbone, conditioned on the instance feature extracted from the data sample to their left ($\mathbf{x}_i$) and their corresponding class.

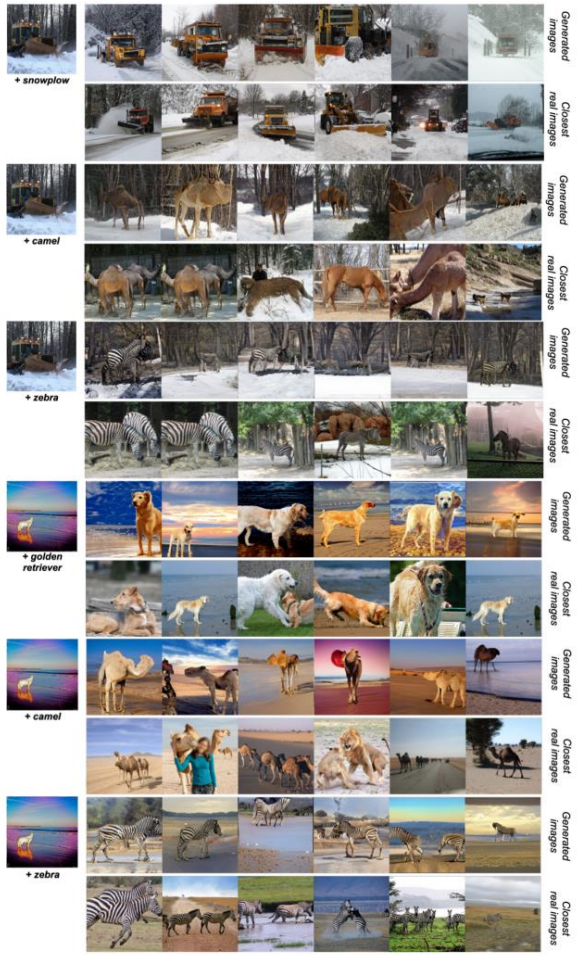


Figure 9: Generated 256×256 images with a class-conditional IC-GAN (BigGAN backbone) trained on ImageNet. Next to each data sample $\mathbf{x}_i$, used to obtain the instance features $\mathbf{h}_i = f_\theta(\mathbf{x}_i)$, we find generated images sampled from IC-GAN using $\mathbf{h}_i$ and six sampled noise vectors. Below the generated images, the closest image in the ImageNet training set are shown (Cosine similarity in the feature space of f_θ).

Additional Qualitative Results

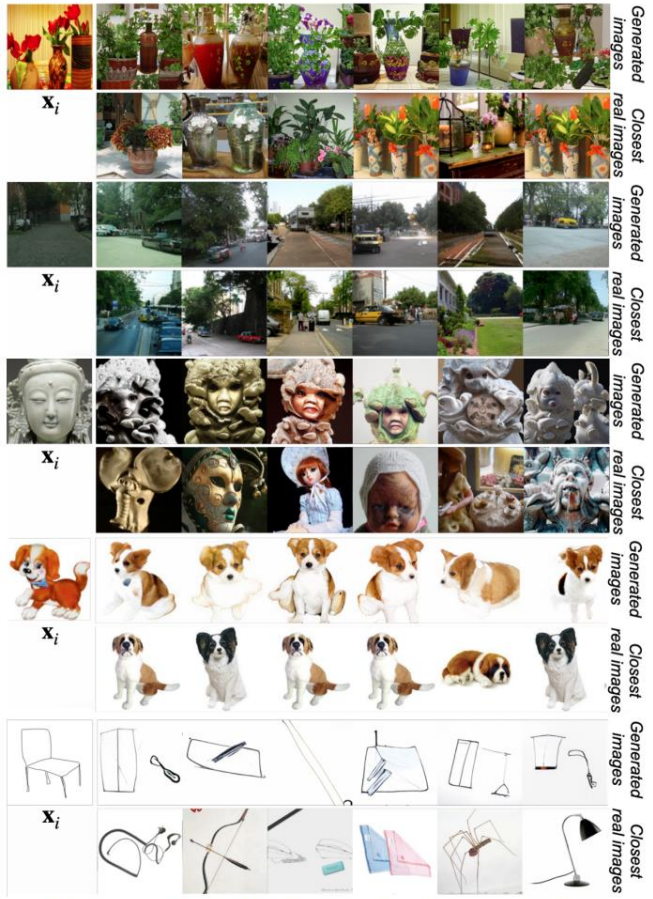


Figure 10: Qualitative off-the-shelf transfer results in 256×256, using an IC-GAN trained on unlabeled ImageNet and conditioning on unseen instances from other datasets. The instances come from the following datasets (top to bottom): COCO-Stuff, Cityscapes, MetFaces, PACS (cartoons), Sketches. Next to each data sample $\mathbf{x}_i$, used to obtain the instance features $\mathbf{h}_i = f_\theta(\mathbf{x}_i)$, generated images conditioning on $\mathbf{h}_i$ are displayed. Immediately below each generated image, the closest image in the ImageNet training set is displayed (Euclidean distance in the feature space of f_θ).

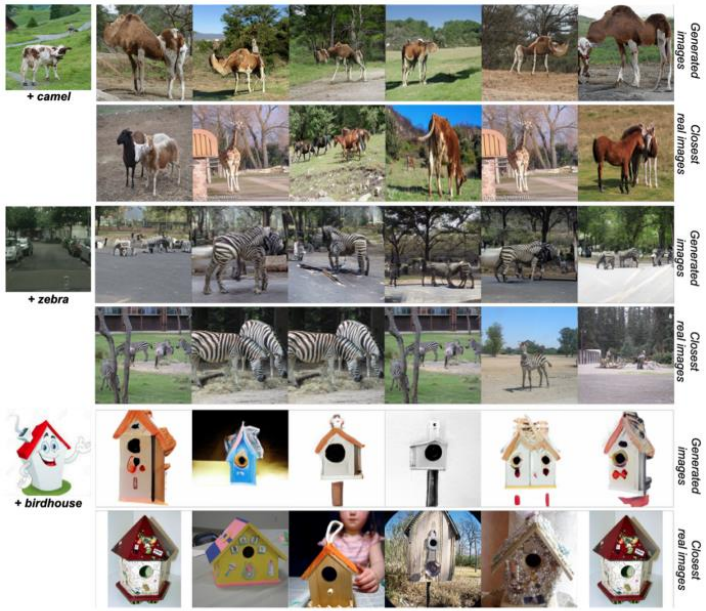


Figure 11: Qualitative off-the-shelf transfer results in 256×256, using a class-conditional IC-GAN trained on ImageNet and conditioning on unseen instances from other datasets and a class label. The instances come from the following datasets (top to bottom): COCO-Stuff, Cityscapes, PACS (cartoons). On the left, a data sample $\mathbf{x}_i$ is depicted, used to obtain the instance features $\mathbf{h}_i = f_\theta(\mathbf{x}_i)$. Next to the data samples, generated images conditioning on $\mathbf{h}_i$ and a class label (under the data samples) are displayed. Just below the generated images, the closest image in the ImageNet training set are shown (Euclidean distance in the feature space of f_θ).

Limitations

- IC-GAN requires storing training instances to use the model, e.g., 1000 instances for IN.
- Feature extractor and generator are not trained jointly.
- Transfer potential degrades in out-of-domain datasets such as very different from IN.