FineGAN: Unsupervised Hierarchical Disentanglement for Fine-Grained Object Generation and Discovery

Krishna Kumar Singh* Utkarsh Ojha* Yong Jae Lee

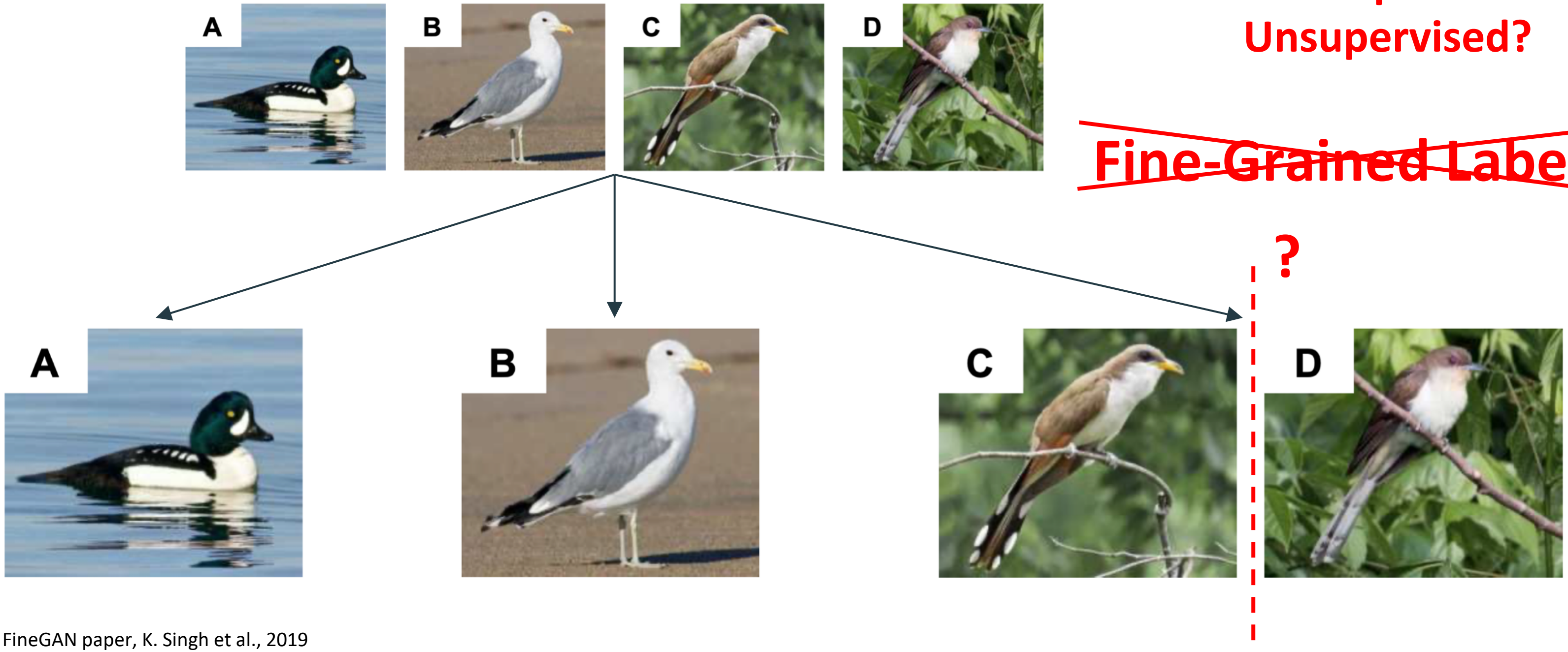
Presenters: Haitian Chen, Haotian Liu, Youzhi Tian

Problem Statement

Image Grouping Problem

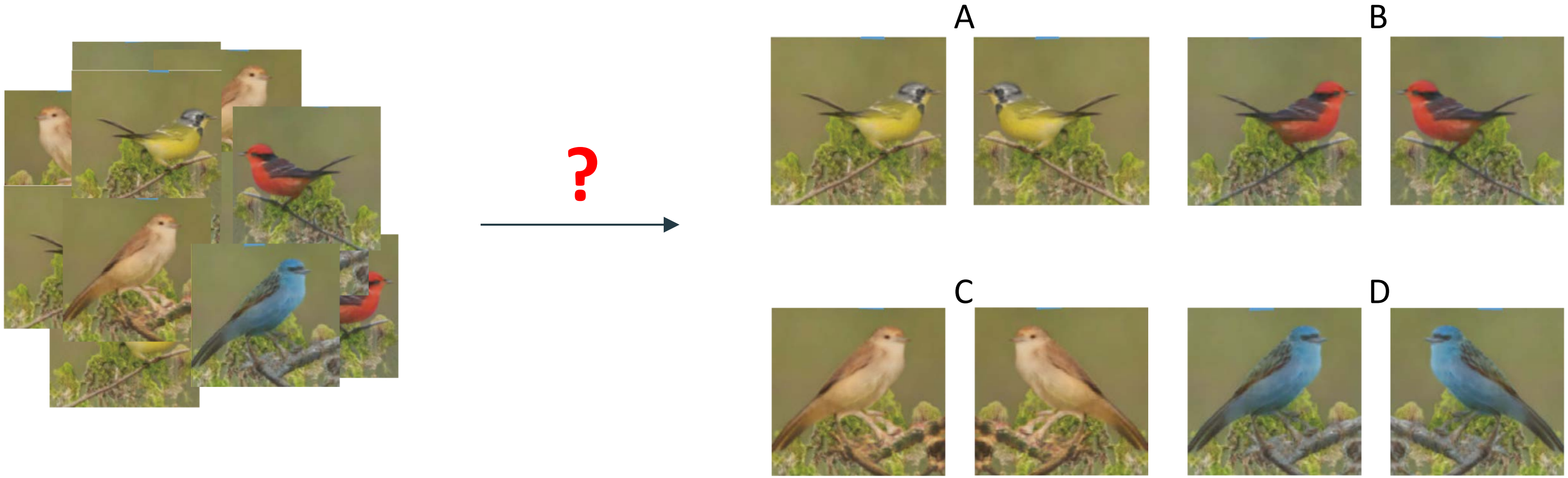
Too expensive!
Unsupervised?

~~Fine-Grained Labels!~~



Images from FineGAN paper, K. Singh et al., 2019

Auto Image Generation



Images from FineGAN paper, K. Singh et al., 2019

Hierarchical Image Generation

Diagram from FineGAN paper, K. Singh et al., 2019

GAN: Generative Adversarial Network

- Generative Model: describes how data is generated, in terms of a probabilistic model.

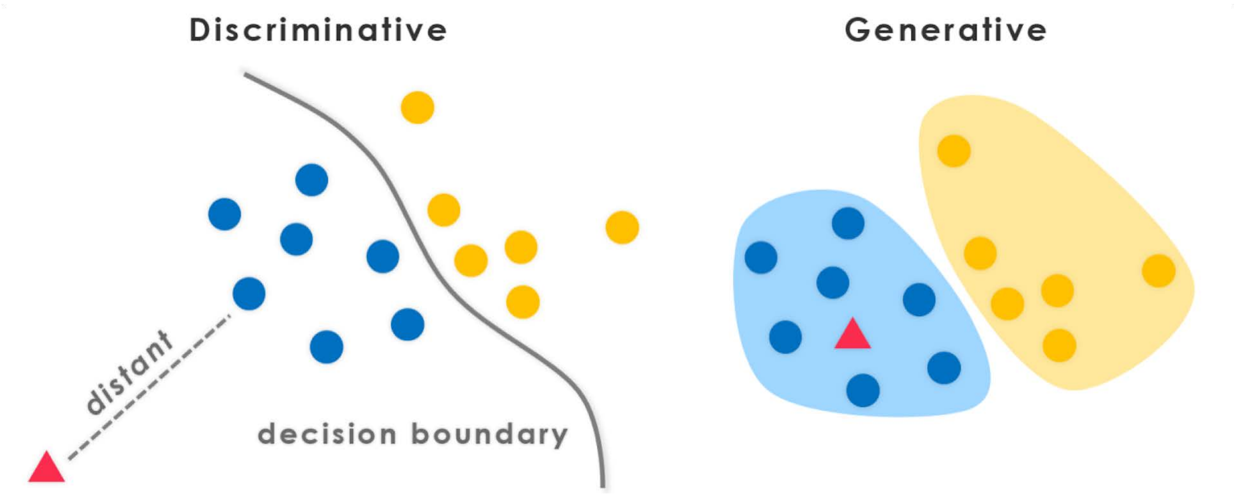


Image credit: Ilya Verenich et al.

GAN: an implicit generative model

- Adversarial Model: Discriminator vs Generator

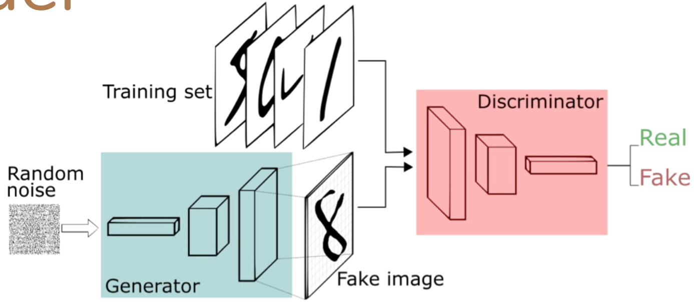
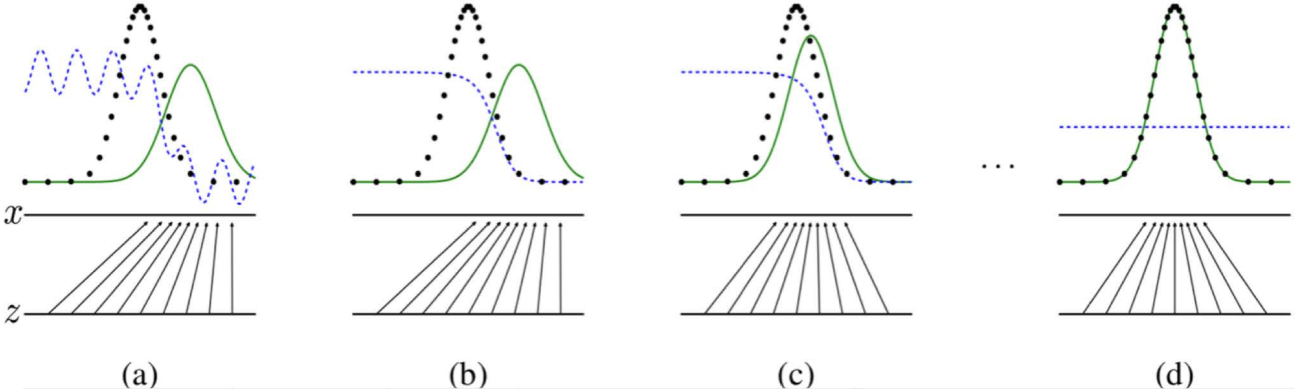


Image credit: Thalles Silva



$G(z)$

$z \sim P_{\text{noise}}(z)$

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim P_{\text{data}}}[\log D(x)] + \mathbb{E}_{z\sim\text{noise}}[\log(1 - D(G(z)))]$$

FineGAN: Fine-Grained GAN

Approach:

- Hierarchically generating and stitching images together

- Disentangle factors / Parent and child latent code
  (vector of latent space of feature)

Related Works

Related work

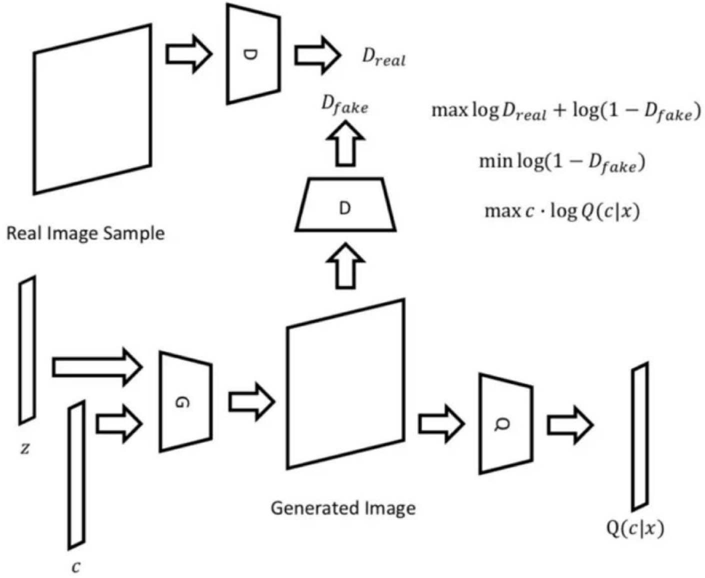
InfoGAN*: Mutual Information between latent codes and images

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X)$$

$$\min_G \max_D V_I(D,G) = V(D,G) - \lambda I(c; G(z,c))$$

z is random noise and **c** is latent code (with label information)

Maximize MI between latent codes and generated images.



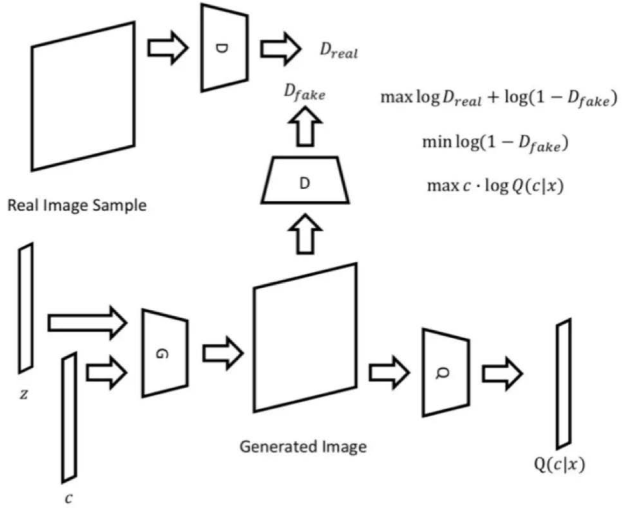
*InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

Related work

InfoGAN*: Mutual Information between latent code

$I(c; G(z,c))$ hard to maximize directly as it requires access to the posterior of $P(c|x)$

$$
\begin{aligned}
I(c; G(z,c)) &= H(c) - H(c|G(z,c)) \\
&= \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log P(c'|x)]] + H(c) \\
&= \mathbb{E}_{x\sim G(z,c)}[\underbrace{D_{\mathrm{KL}}(P(\cdot|x) \parallel Q(\cdot|x))}_{\geq 0} + \mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c) \\
&\geq \mathbb{E}_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c)
\end{aligned}
$$



Find a lower bound of it by defining an auxiliary distribution Q(c|x) to approximate P(c|x).

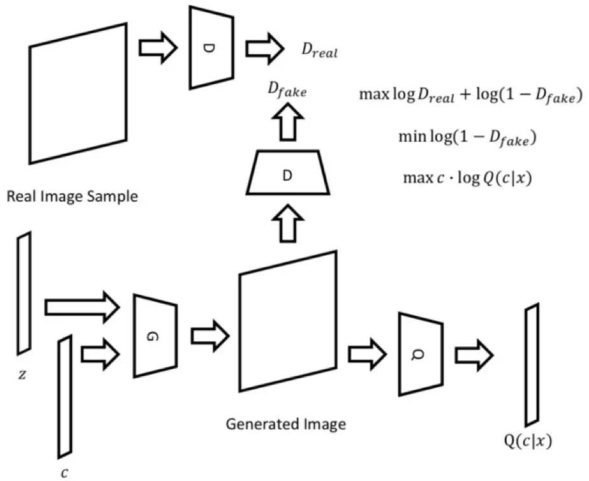
*InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

Related work

InfoGAN*: Mutual Information between latent code

$$L_I(G,Q) = E_{x\sim G(z,c)}[\mathbb{E}_{c'\sim P(c|x)}[\log Q(c'|x)]] + H(c)$$

$$\min_{G,Q}\max_{D} V_{\text{InfoGAN}}(D,G,Q) = V(D,G) - \lambda L_I(G,Q)$$



*InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets

Related work

Fine-grained category recognition

- involves classifying subordinate categories within entry-level categories

Visual object discovery and clustering

- unsupervised object discovery

Disentangled representation learning

- InfoGAN

GANs and Stagewise image generation

- Unconditional GANs

Approach

Big Picture

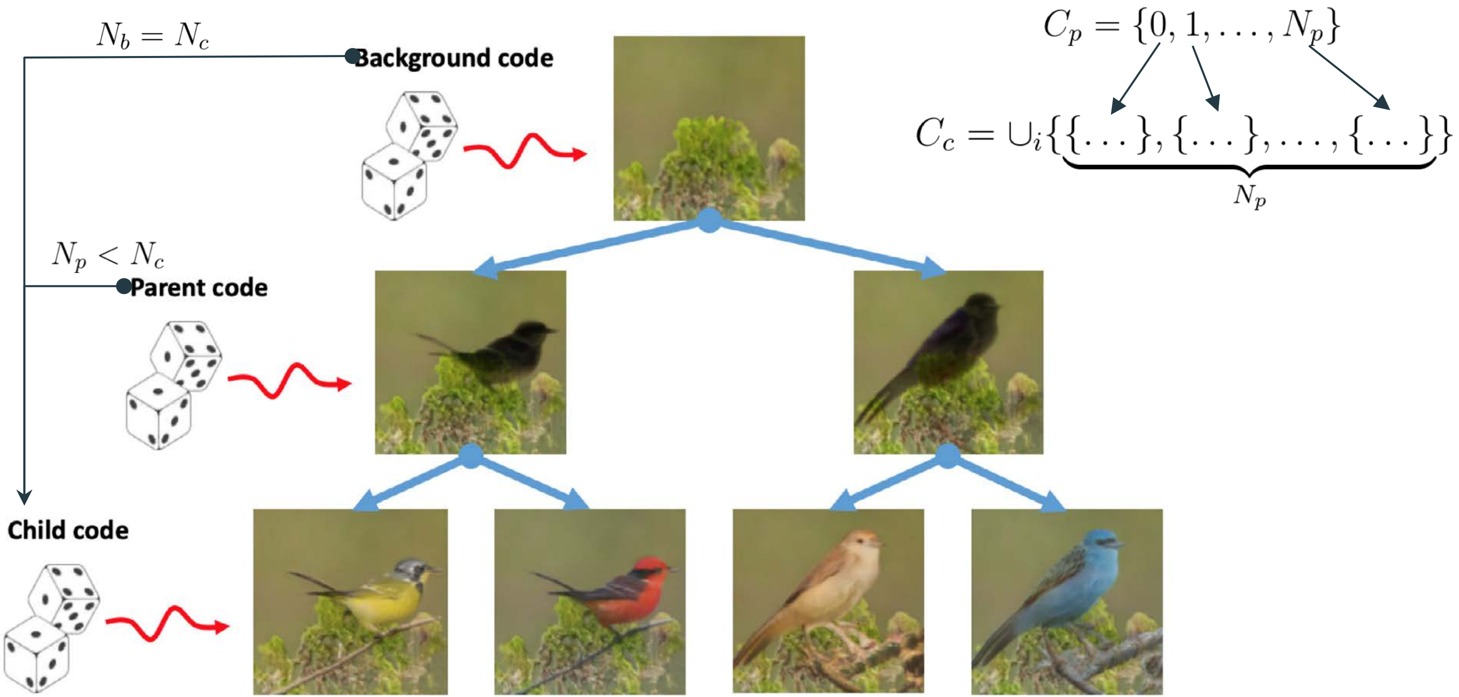


Diagram from FineGAN paper, K. Singh et al., 2019

Architecture Overview

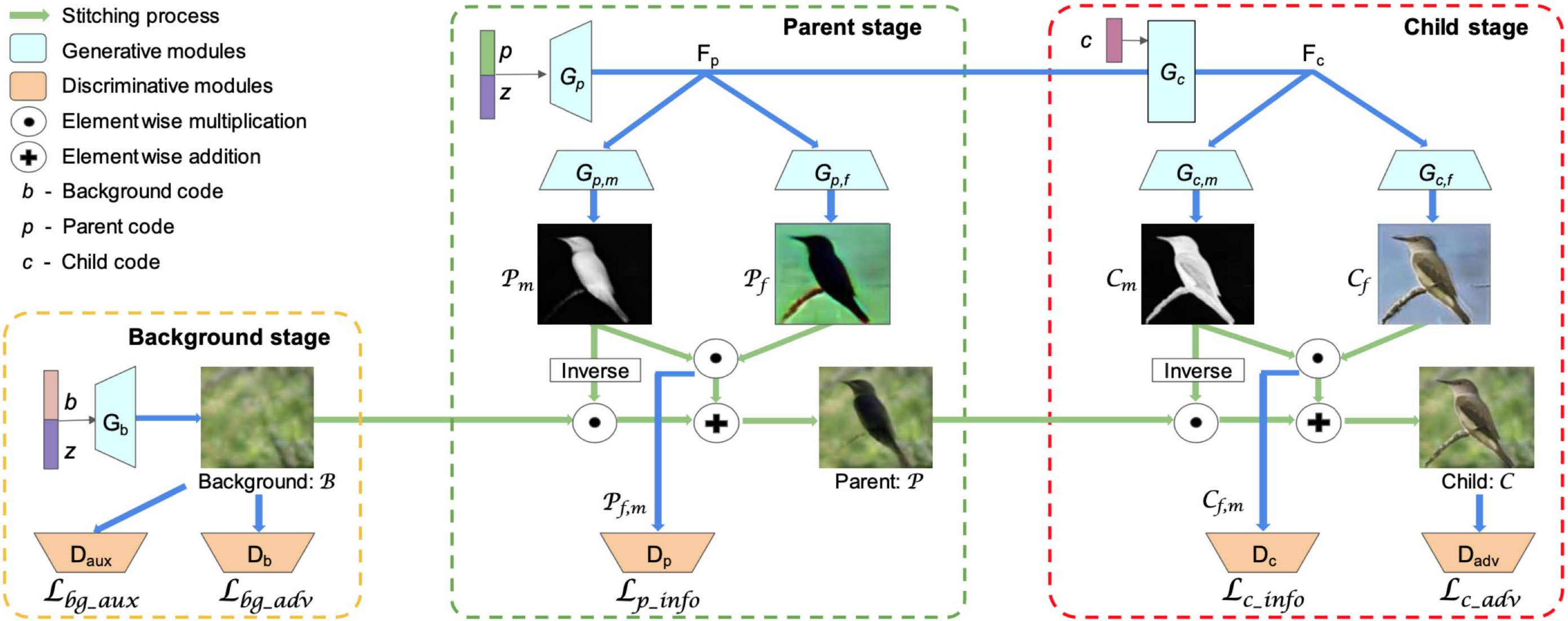


Diagram from FineGAN paper, K. Singh et al., 2019

Architecture Overview

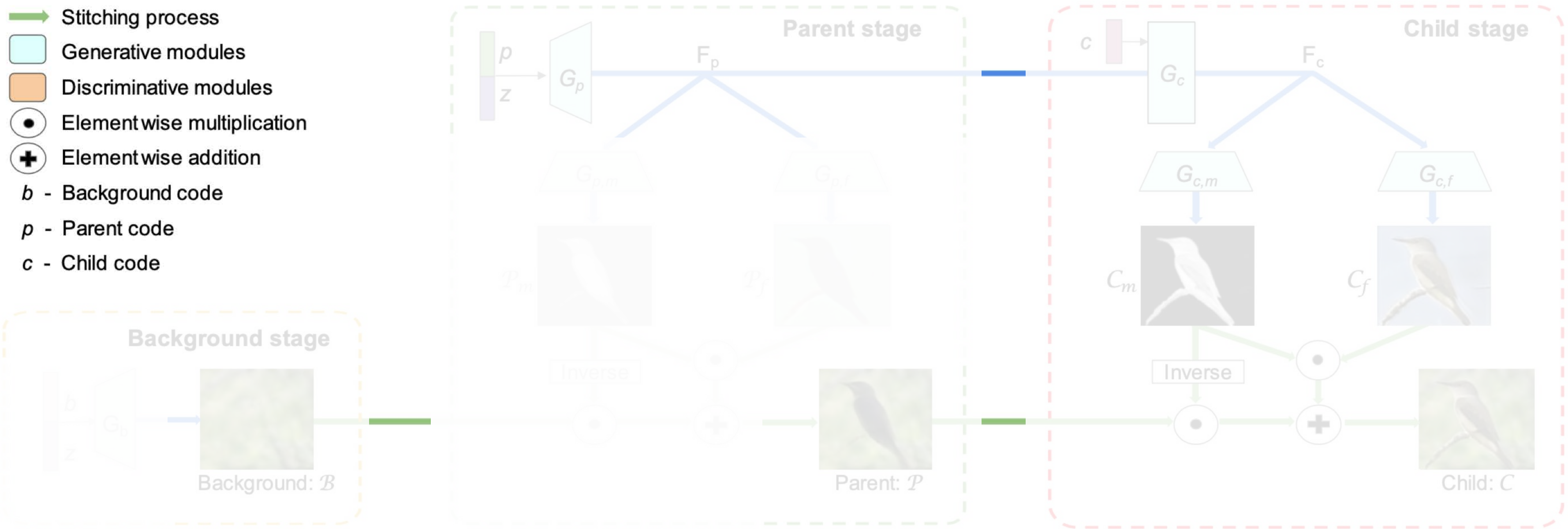


Diagram from FineGAN paper, K. Singh et al., 2019

Training Losses

$$\mathcal{L}_{bg\_adv} = \min_{G_b} \max_{D_b} \mathbb{E}_x[\log(D_b(x))] + \mathbb{E}_{z,b}[\log(1 - D_b(G_b(z,b)))]$$

$$\mathcal{L}_{c\_adv} = \min \max_{D_{adv}} \mathbb{E}_x[\log(D_{adv}(x))] + \mathbb{E}_{z,b,p,c}[\log(1 - D_{adv}(C))]$$

$$\mathcal{L}_p = \mathcal{L}_{p\_info} = \max_{D_p, G_{p,f}, G_{p,m}} \mathbb{E}_{z,p}[\log D_p(p|\mathcal{P}_{f,m})]$$

$$\mathcal{L}_{c\_info} = \max_{D_c, G_{c,f}, G_{c,m}} \mathbb{E}_{z,p,c}[\log D_c(c|\mathcal{C}_{f,m})]$$

$$\mathcal{L}_{bg\_aux} = \min_{G_b} \mathbb{E}_{z,b}[\log(1 - D_{aux}(G_b(z,b)))]$$

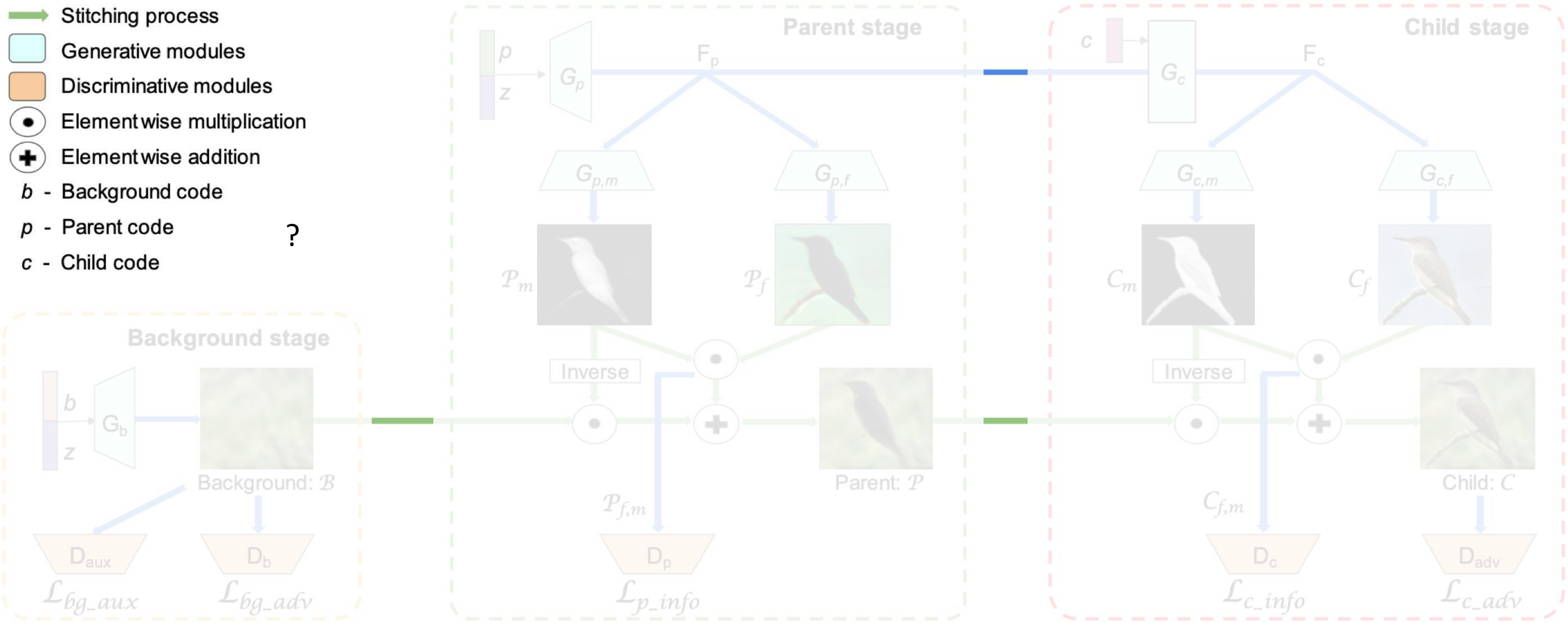


Diagram from FineGAN paper, K. Singh et al., 2019

Training Classifier

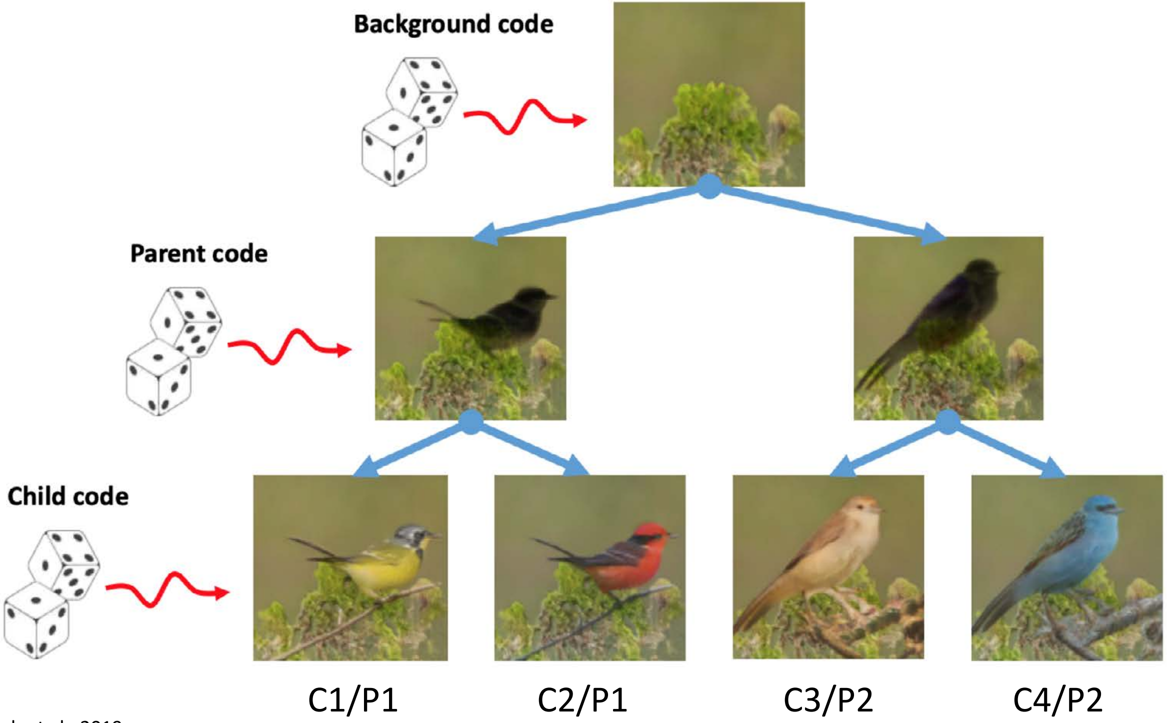


Diagram from FineGAN paper, K. Singh et al., 2019

Experiments

Experimental setup and results

Dataset:

(1) CUB: 200 bird classes (11788 images).
(2) Stanford Dogs: 120 dog classes (training data 12000 images).
(3) Stanford Cars: 196 car classes (training data 8144 images).

Number of parents and children:

(1) CUB: N_p = 20 N_c = 200
(2) Stanford Dogs: N_p = 12 N_c = 120
(3) Stanford Cars: N_p = 20 N_c = 196

Task:

(1) Fine-grained image generation
(2) Fine-grained object category discovery

Fine-grained image generation

Baselines:

(1) **Simple-GAN**: generates a final image in one shot without the parent and background stages.
(2) **InfoGAN**: same as Simple-GAN but with additional $\mathcal{L}_{c\_info}$.
(3) **LR-GAN**: it also generates an image stagewise but it stage only consists of foreground and background.
(4) **StackGAN-v2**: its unconditional version generates images at multiple scales with $\mathcal{L}_{c\_adv}$ at each scale.

Evaluation:

(1) Quantitative evaluation of image generation.
(2) Qualitative evaluation of image generation.

Quantitative evaluation of image generation

Metric:

(1) Inception Score (**IS**).
(2) Frechet Inception Distance (**FID**).

Results:

	IS			FID		
	Birds	Dogs	Cars	Birds	Dogs	Cars
Simple-GAN	31.85 ± 0.17	6.75 ± 0.07	20.92 ± 0.14	16.69	261.85	33.35
InfoGAN [9]	47.32 ± 0.77	43.16 ± 0.42	28.62 ± 0.44	13.20	29.34	17.63
LR-GAN [50]	13.50 ± 0.20	10.22 ± 0.21	5.25 ± 0.05	34.91	54.91	88.80
StackGANv2 [55]	43.47 ± 0.74	37.29 ± 0.56	$\mathbf{33.69 \pm 0.44}$	13.60	31.39	16.28
FineGAN (ours)	$\mathbf{52.53 \pm 0.45}$	$\mathbf{46.92 \pm 0.61}$	32.62 ± 0.37	**11.25**	**25.66**	**16.03**

Table 1. Inception Score (higher is better) and FID (lower is better). FineGAN consistently generates diverse and real images that compare favorably to those of state-of-the-art baselines.

Quantitative evaluation of image generation

How sensitive is FineGAN to the number of parents:

	N_p=20	N_p=10	N_p=40	N_p=5	N_p=mixed
Inception Score (CUB)	**52.53**	52.11	49.62	46.68	51.83

Table 2. Varying number of parent codes N_p, with number of children N_c fixed to 200. FineGAN is robust to a wide range of N_p.

With variable number of children per parent (Np=mixed: 6 parents with 5 children, 3 parents with 20 children, and 11 parents with 10 children), IS remains high, which shows there is no need to have the same number of children for each parent.

Qualitative evaluation of image generation

(1) Image generation process.

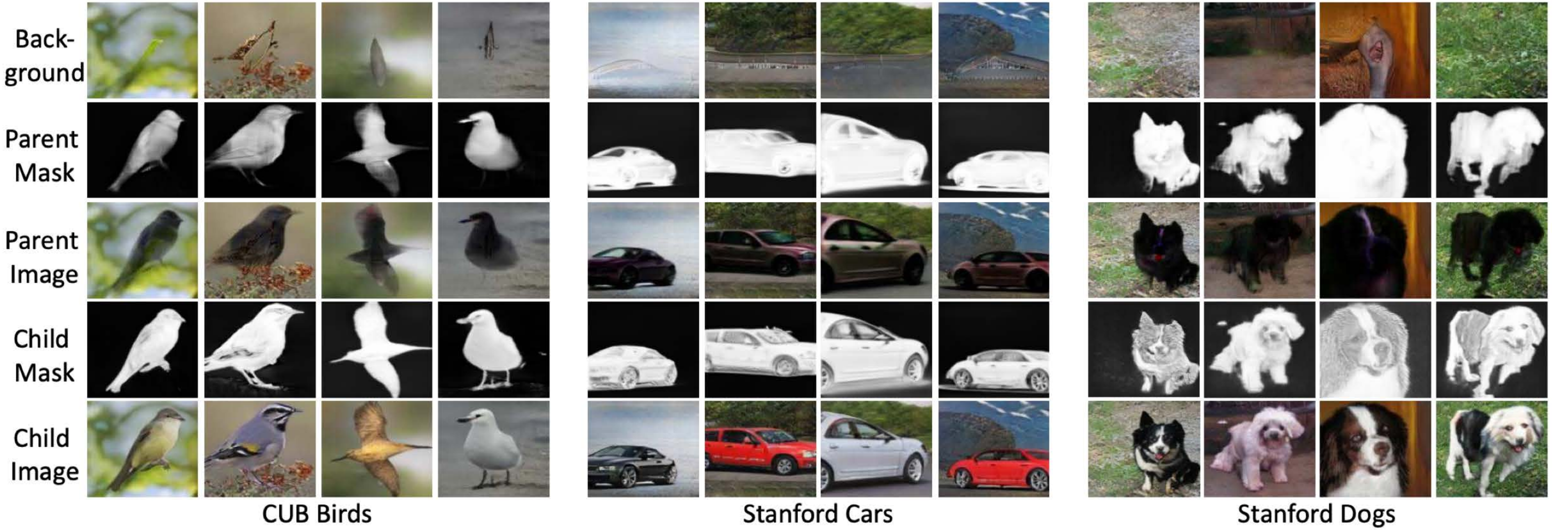


Figure 3. **FineGAN's stagewise image generation.** Background stage generates a background which is retained over the child and parent stages. Parent stage generates a hollow image with only the object's shape, and child stage fills in the appearance to complete the image.

Qualitative evaluation of image generation

(2) Disentanglement of factors of variation.

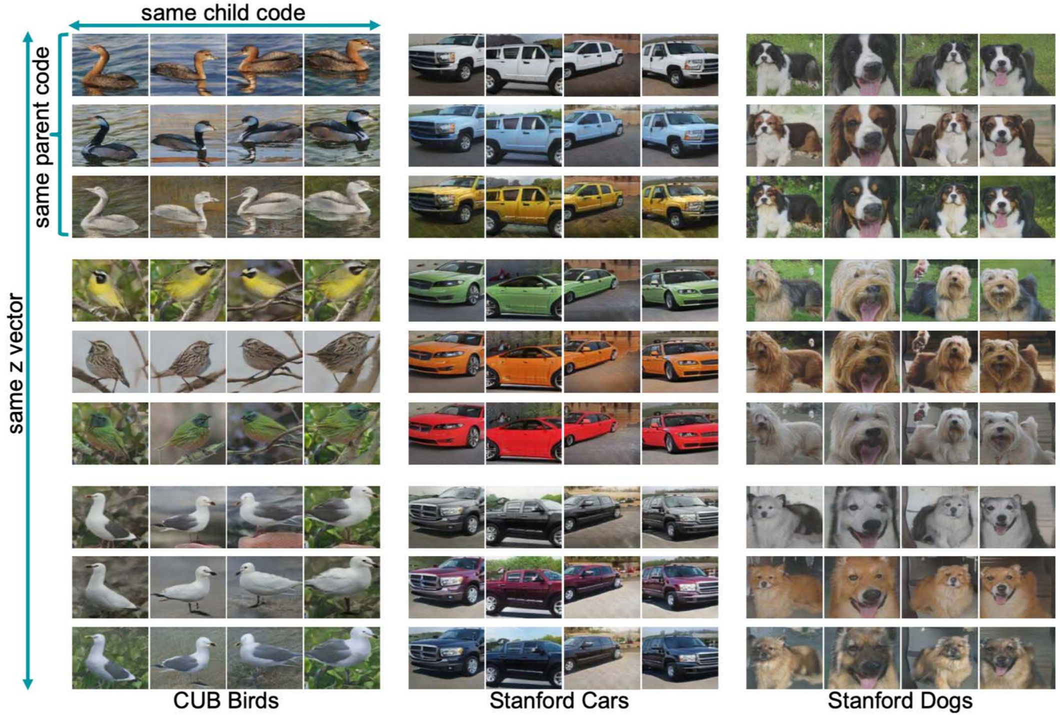


Figure 4. **Varying p vs. c vs. z.** Every three rows correspond to the same parent code p and each row has a different child code c. For the same parent, the object's shape remains consistent while the appearance changes with different child codes. For the same child, the appearance remains consistent. Each column has the same random vector z – we see that it controls the object's pose and position.

Qualitative evaluation of image generation

(3) Disentanglement of parent vs. child.

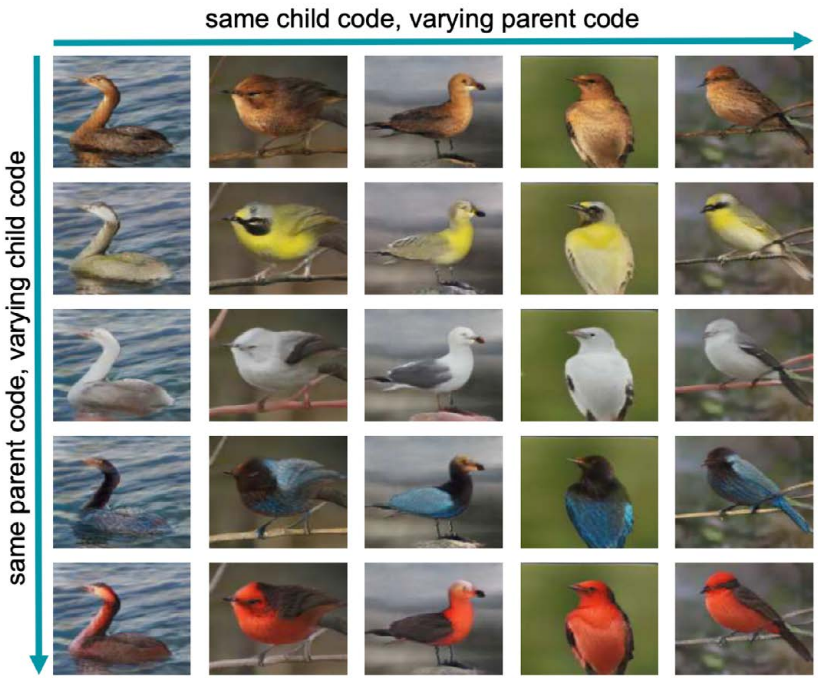


Figure 5. **Disentanglement of parent vs. child codes.** Shape is retained over the column, appearance is retained over the row.

Qualitative evaluation of image generation

(4) Disentanglement of background vs. foreground


(a) Fixed b, varying p and c


(b) Fixed p and c, varying b

Qualitative evaluation of image generation

(5) Comparison with InfoGAN.

Figure 6. **InfoGAN results.** Images in each group have same child code. The birds are the same, but so are their backgrounds. This strongly suggests InfoGAN takes background into consideration when categorizing the images. In contrast, FineGAN's generated images (Fig. 4) for same c show reasonable variety in background.

Fine-grained object category discovery

Baselines:

(1) JULE
(2) DEPICT
(3) JULE-ResNet-50
(4) DEPICT-Large

Metric:

(1) Normalized Mutual Information (**NMI**)
(2) Accuracy (of best mapping between cluster assignments and true labels)

Fine-grained object category discovery

	NMI			Accuracy		
	Birds	Dogs	Cars	Birds	Dogs	Cars
JULE [51]	0.204	0.142	0.232	0.045	0.043	0.046
JULE-ResNet-50 [51]	0.203	0.148	0.237	0.044	0.044	0.050
DEPICT [15]	0.290	0.182	0.329	0.061	0.052	0.063
DEPICT-Large [15]	0.297	0.183	0.330	0.061	0.054	0.062
Ours	**0.403**	**0.233**	**0.354**	**0.126**	**0.079**	**0.078**

Table 3. Our approach outperforms existing clustering methods.

Strengths and weakness

Strengths:

(1) Accurately disentangle background, object shape, and object appearance.

(2) Generate realistic and diverse images.

(3) Produce fine-grained clusters that are significantly more accurate than those of state-of- the-art unsupervised clustering approaches.

Weakness:

(1) The number of children are hyperparameters that a user must set, which can be difficult when the true number of categories is unknown (a problem common to most unsupervised grouping methods).

(2) The latent modes of variation that FineGAN discovers may not necessarily correspond to those defined/annotated by a human.

(3) we are far behind fully-supervised fine-grained recognition methods.

Applications:

(1) Style transfer
(2) Image clustering

Contributions

Introduces an unsupervised model that learns to hierarchically generate the background, shape, and appearance of fine-grained object categories.

Learns disentangled representation to cluster real images for unsupervised fine-grained object category discovery.