ComGAN: Unsupervised Disentanglement and Segmentation via Image Composition

Rui Ding, Kehua Guo, Xiangyuan Zhu, Zheng Wu, and Liwei Wang

Central South University

NEURAL INFORMATION PROCESSING SYSTEMS

Background

- **Image composition** can be regarded as combining multiple visual areas to construct a realistic image.
  - This kind of work commonly requires the following assumption:

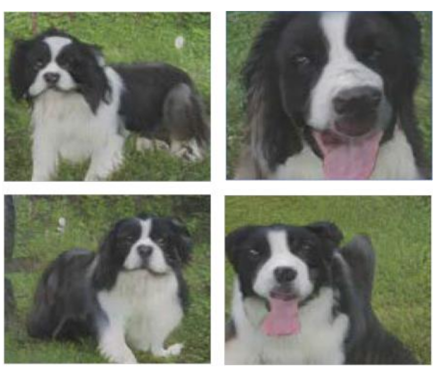
Assumption 1 An image x taken from the world is typically composed of foreground x_f and background x_b, which can be decomposed by the following equation:

$$x = x_f \odot x_m + x_f \odot (1 - x_m),$$

where x_m is the mask, and the $\odot$ denotes element-wise multiplication operator.

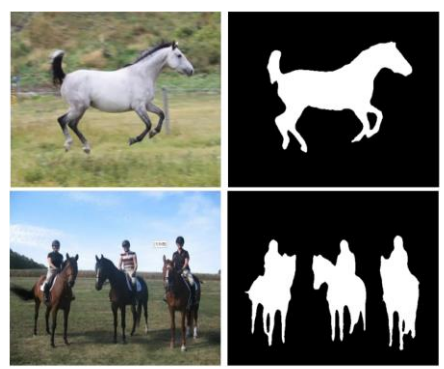
- **Application:**

Image Disentanglement



Images credit: FineGAN[1]

Object Segmentation



Images credit: Labels4Free[2]

Clustering

Images credit: C3-GAN[3]

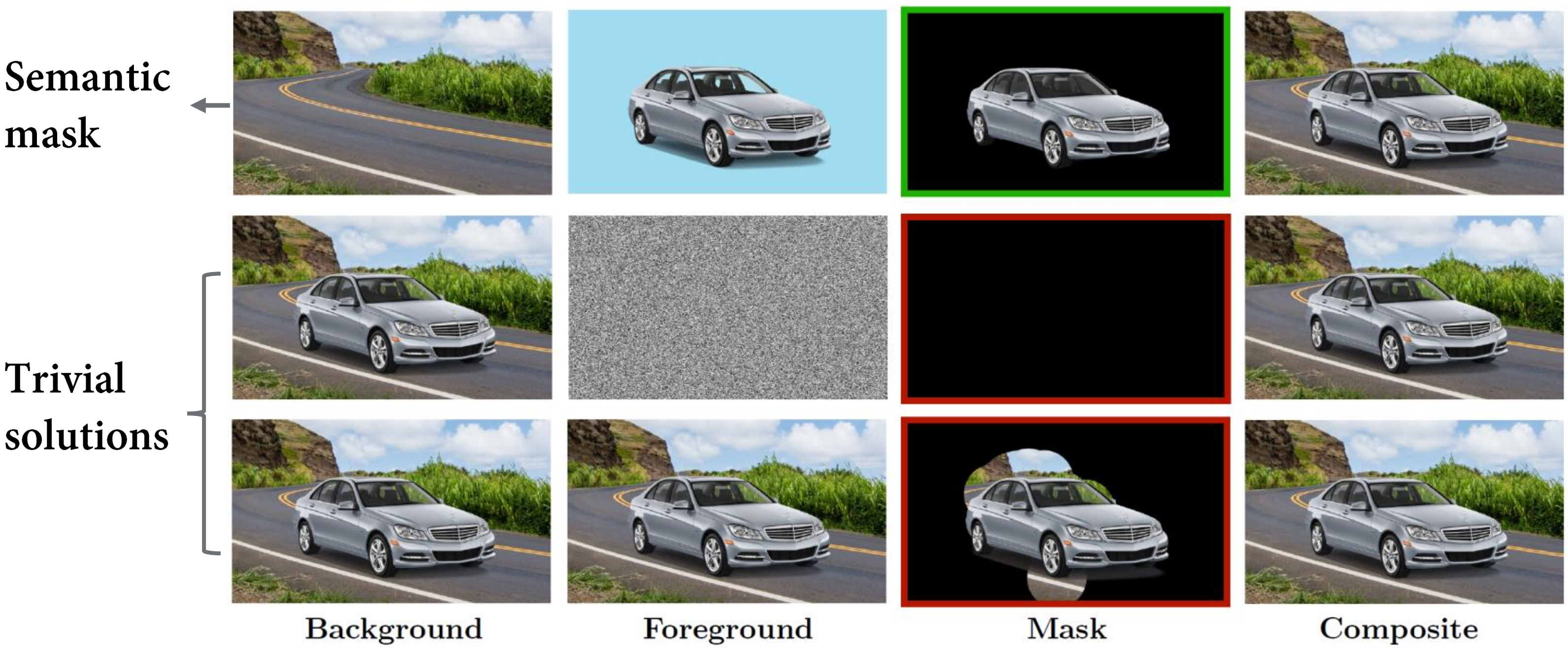
[1] Singh, Krishna Kumar, et al., CVPR, 2019; [2] Abdal, Rameen, et al., ICCV, 2021; [3] Kim, Yunji, et al., ICLR 2022.

Background

Assumption 1 works so well. Does it have a negative effect?

- The compositional generation process is often accompanied by trivial solutions.
  - **Trivial solutions** can be considered meaningless masks generated by models.

Images credit:PerturbGAN [4]



[4] Adam Bielski, et al., NeurIPS, 2019.

Motivation

◆ **Previous methods:**

- Add supervised information to the model.
- Design clever regularization and fine-tune the parameters.

※ **Our goal is:**

- **finding the source of trivial solutions, and**
- **designing a model that can**
  - Solve this issue from other perspectives, and
  - Effective extension to relevant applications.

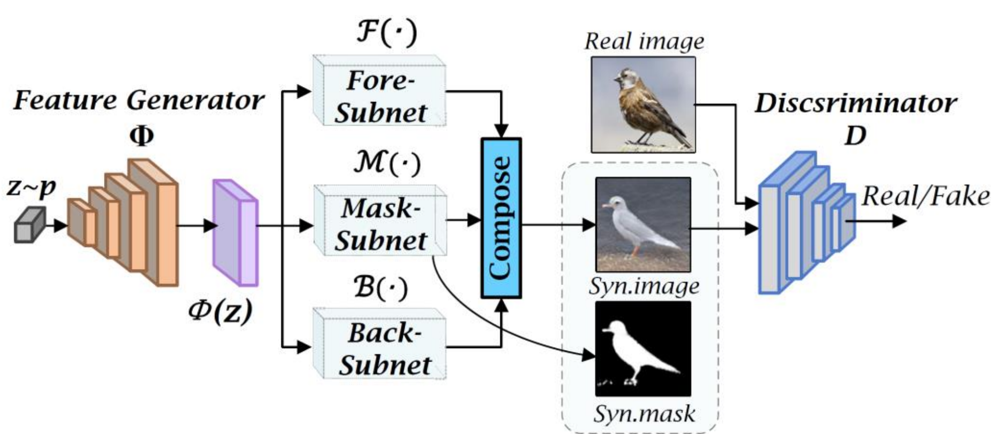
Method

- **The source of trivial solutions:**

> ***Lemma 1*** Let L_{all} be the overall loss and x_m be the synthesized mask. Consider a model that composes images utilizing Assumption 1. There exist vanishing gradients on the mask if and only if the model converges to two kinds of trivial solutions.

Key idea: Keep the gradients positive from the perspective of architecture.

- **Framework of ComGAN:**

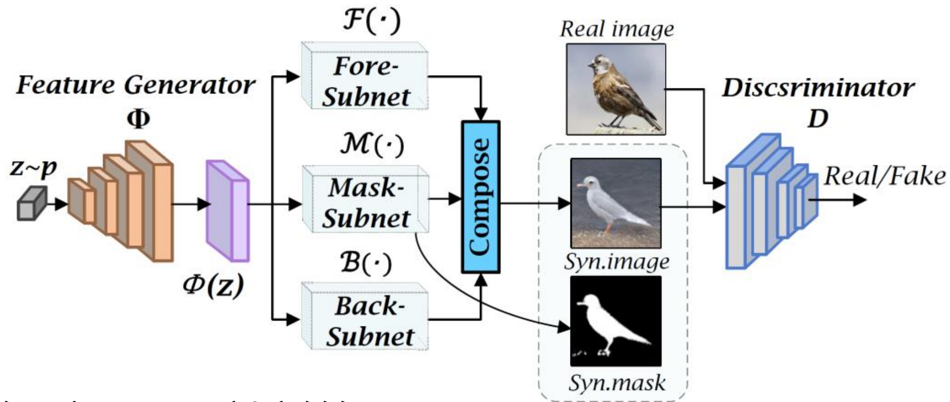


- The learning objective is as follows:

$$L_{all} = \min_{\Phi,\mathcal{F},\mathcal{B},\mathcal{M}} \max_{D} \mathcal{L}_D^{adv} + \min_{\Phi,\mathcal{M}} \beta \mathcal{L}_{\text{binary}}$$

Method

※ Synthesize images and high semantic masks in an unsupervised manner.



➢ The composited image can be written as:

$$\bar{x} = \mathcal{F}\,(\phi(z)) \odot \mathcal{M}(\phi(z)) \;+\; \mathcal{B}(\phi(z)) \odot (1 \;-\; \mathcal{M}(\phi(z)))$$

◆ **Advantages of ComGAN:**

➢ This form generalizes two typical image compositional generation methods:

Model $\prod_1$ $\mathcal{B}\left(\phi(z)\right) = \mathcal{B}(z)$: FineGAN [1] , Labels4Free[2], etc. can be formulated as this form;

Model $\prod_2$ $\phi(z) = z$: PerturbGAN [4] , CGN [5], etc. can be formulated as this form.

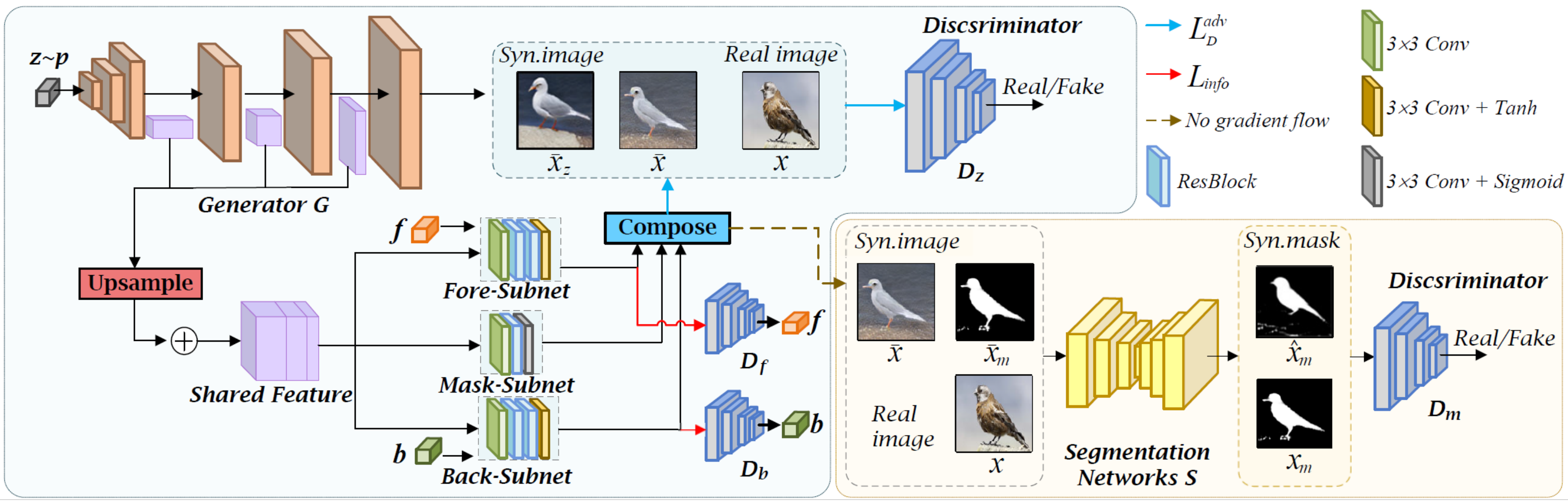
➢ In this paper, Theorem 1 shows that there exists a lower bound on the gradient norms of the mask generator in ComGAN.

[5] Axel Sauer, et al., ICLR, 2021.

Method

Can the advantages of ComGAN be reflected in relevant applications?

- **DS-ComGAN,** the ComGAN-based variant, achieves image disentanglement and object segmentation in a fully unsupervised manner.

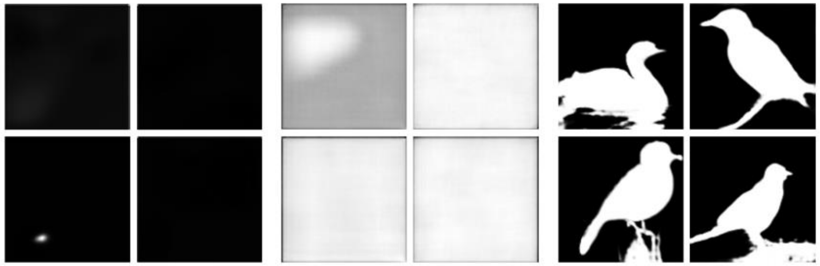


- The learning objective is as follows:

$$L_{all} = \overbrace{\max_{D_f, D_b} L_{info} + \min_{G,\mathcal{F},\mathcal{B},\mathcal{M}} \max_{D_z} \mathcal{L}_{D_z}^{adv}}^{\text{objective for image disentanglement task}} + \underbrace{\min_{\mathcal{S}} \max_{D_m} \mathcal{L}_{D_m}^{adv} + \min_{\mathcal{S}} \lambda \mathcal{L}_{\text{cons}}}_{\text{objective for object segmentation task}}$$

Experiments #1: Com-GAN

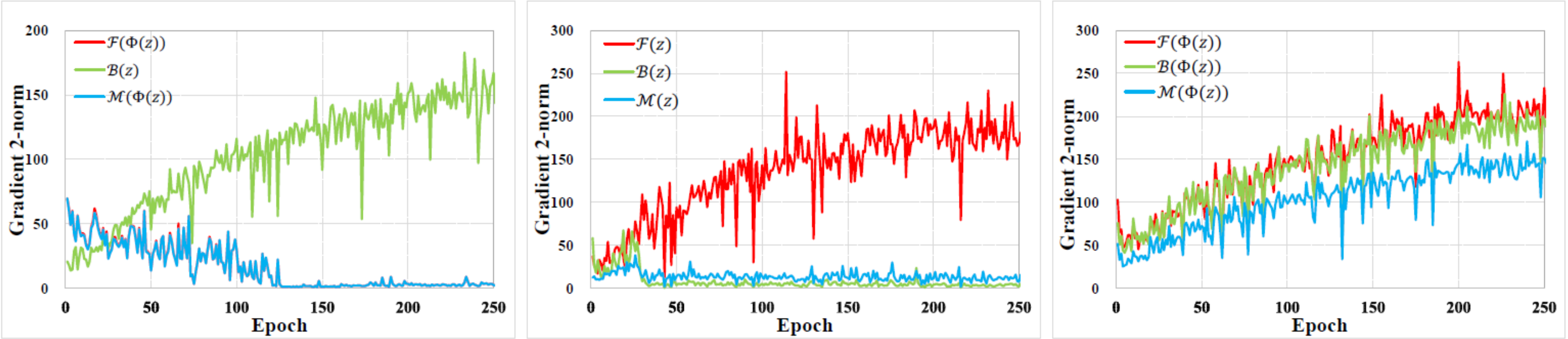
※ To show that **ComGAN solves trivial solutions**, the comparison results on CUB for the synthesized mask and the gradient norms are presented.



(a) Model Π_1 (b) Model Π_2 (c) ComGAN

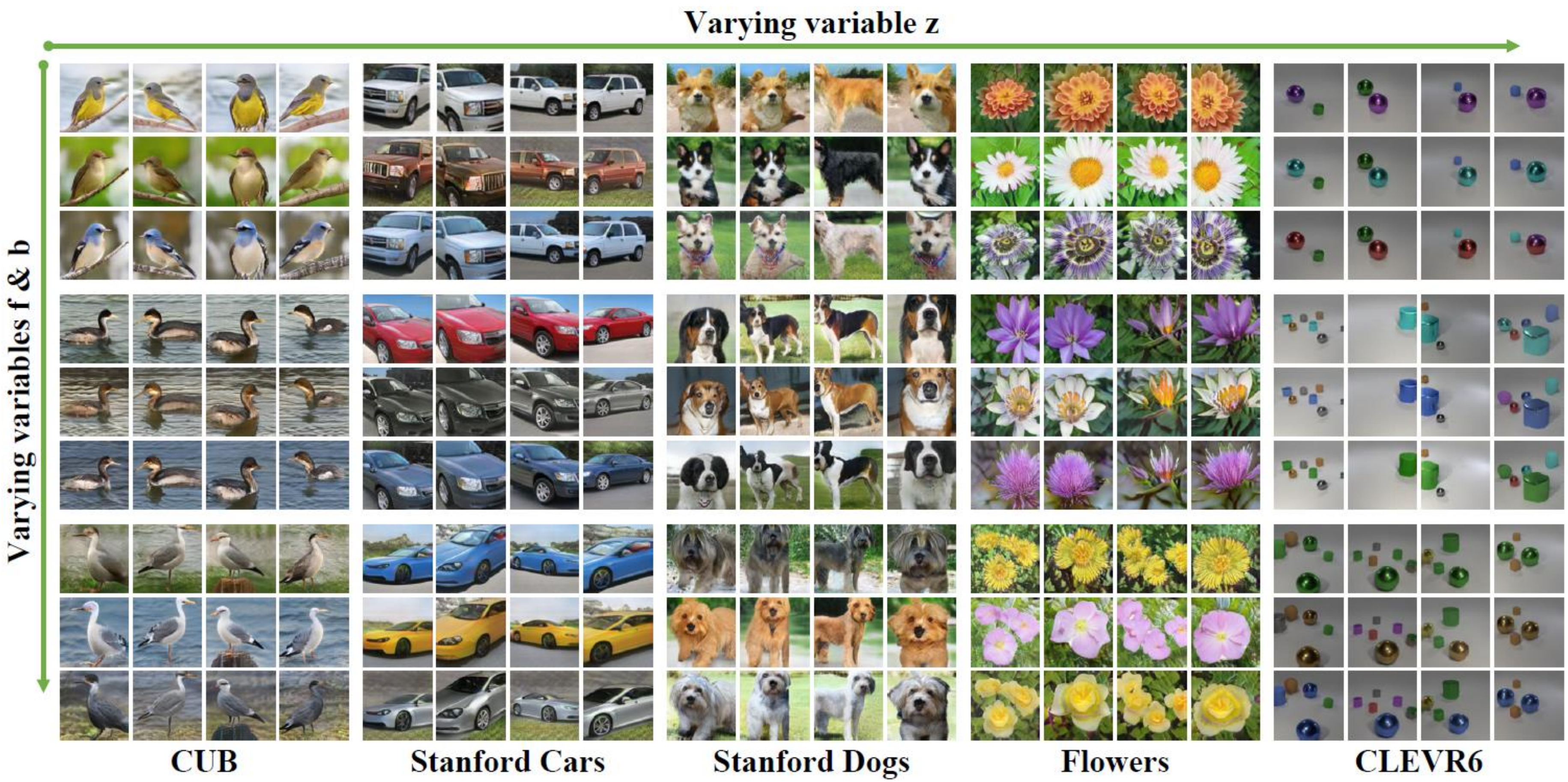
FineGAN as a typical network of model $\prod_1$.

PerturbGAN (Replace StyleGAN with simpleGAN) **as a typical network of** model $\prod_2$.



(a) Model Π_1 (b) Model Π_2 (c) ComGAN

Experiments #2: DS-ComGAN



Experiments #2: DS-ComGAN

Methods	Sup.	CUB FID ↓	CUB IS↑	FS-100 FID ↓	FS-100 IS↑	Stanford-Cars FID ↓	Stanford-Cars IS↑
Triple-GAN [49]	Semi.	140.94	3.94±0.06	91.05	1.45±0.03	114.12	2.45±0.06
EnhancedTGAN [40]	Semi.	133.57	4.17±0.06	57.58	1.57±0.02	105.20	2.43±0.05
Triangle-GAN [50]	Semi.	96.42	4.36±0.05	35.49	1.71±0.04	61.44	2.77±0.10
R^3-CGAN [51]	Semi.	88.62	4.43±0.06	25.28	1.73±0.02	44.57	3.05±0.04
SSC-GAN§ [27]	Semi.	20.03	4.68±0.04	20.65	1.82±0.03	39.02	**3.10±0.03**
FineGAN§ [25]	Weak.	46.68	4.62±0.03	24.63	1.76±0.02	45.72	2.85±0.04
MixNMatch§ [31]	Weak.	45.59	4.78±0.08	25.63	1.71±0.05	45.94	2.60±0.05
SN-GAN [52]	Unsup.	160.09	4.21±0.05	41.26	1.66±0.05	53.20	2.80±0.05
DS-ComGAN§	Unsup.	**16.26**	**4.79±0.47**	**20.15**	**1.83±0.32**	**34.17**	2.84±0.12

Table 1: **Image synthesis results for each dataset measured in FID and IS.** DS-ComGAN is compared with the state-of-the-art un(semi-)supervised GAN-based models. § indicates that the models have the ability to achieve image disentanglement.

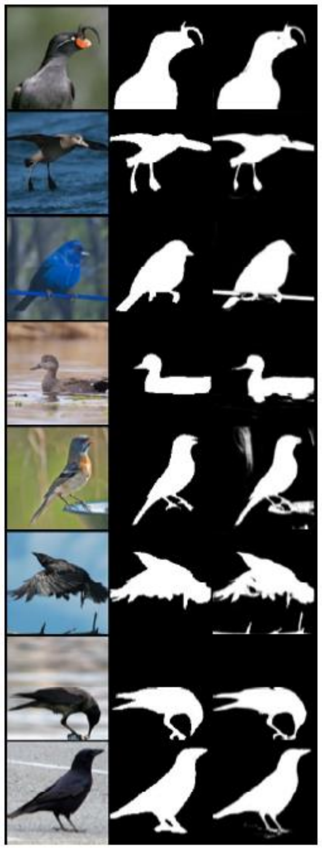
Methods	Single Object: Stanfor-Dogs FID ↓	Single Object: Stanfor-Dogs IS↑	Single Object: Flowers FID ↓	Single Object: Flowers IS↑	Multi-Object: CLEVR6 FID ↓	Multi-Object: CLEVR6 IS↑
SSC-GAN§ [27]	64.26	8.97±0.12	29.09	3.41±0.03	\	\
FineGAN§ [25]	69.52	8.27±0.17	\	\	\	\
MixNMatch§ [31]	68.31	8.32±0.06	\	\	\	\
DS-ComGAN§	**60.84**	**9.17±0.23**	**27.19**	**3.42±0.04**	**77.08**	**2.75±0.05**

Table 2: **Performance of DS-ComGAN on datasets with various attributes.** Noting that the Flowers dataset lacks bounding box annotation, FineGAN and MixNMatch both are unsuitable for this dataset (marked as \ ). CLEVR6 lacks bounding box annotation and labels. Consequently, only our model is suitable for CLEVR6.

※ DS-ComGAN exhibits:

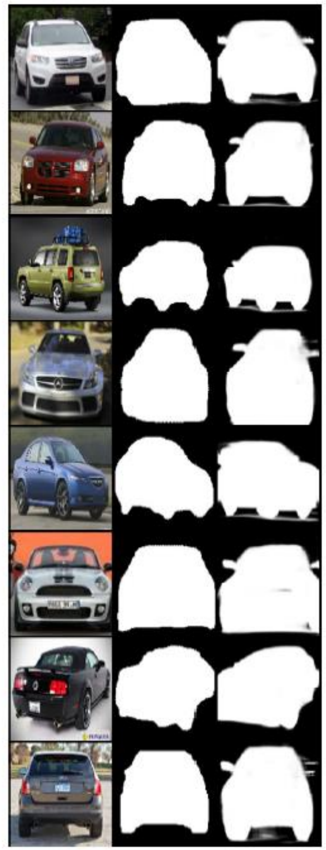
- **Unsupervised image disentanglement** (Varying variables $\boldsymbol{f}$, $\boldsymbol{b}$ & $\boldsymbol{z}$);
- **Better image quality** (FID↑) **and diversity** (IS↓);
- **Robustness to diverse datasets.**

Experiments #2: DS-ComGAN



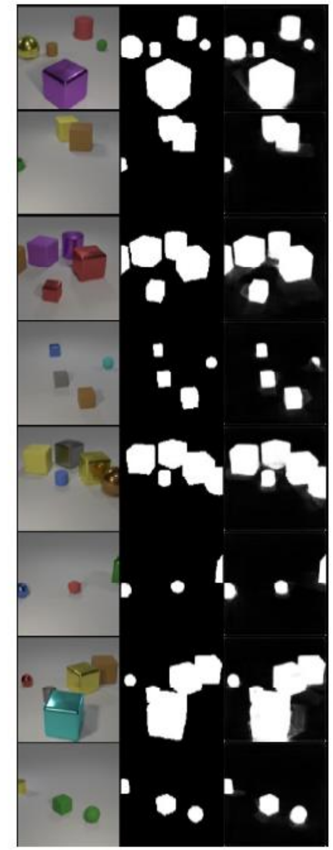
(a) CUB

(b) Stanford-Dogs



(c) Stanford-Cars

(d) Flowers



(e) CLEVR6

Experiments #2: DS-ComGAN

	Single Object								Multi-Object	
	CUB		Stanford-Dogs		Stanford-Cars		Flowers		CLEVR6	
Methods	IoU↑	Dice↑	IoU↑	Dice↑	IoU↑	Dice↑	IoU↑	Dice↑	IoU↑	Dice↑
W-Net [34]	24.8	38.9	47.7	62.1	52.8	67.6	-	-	-	-
GrabCut [53]	30.2	42.7	58.3	70.9	61.3	73.1	69.2	79.1	19.0	30.5
ReDO†⋇ [35]	46.5	60.2	55.7	70.3	52.5	68.6	76.4	-	18.6	31.0
OneGAN⋄⋇ [21]	55.5	69.2	71.0	81.7	71.2	82.6		-	-	-
IODINE† [54]	30.9	44.6	54.4	67.0	51.7	67.3	-	-	19.9	32.4
PerturbGAN [11]	38.0	-	-	-	-	-	-	-	-	-
Slot-Attn.† [55]	35.6	51.5	38.6	55.3	41.3	58.3	-	-	83.6	90.7
IEM+SegNet [36]	55.1	68.7	-	-	-	-	76.8	**84.6**	-	-
DRC [44]	56.4	70.9	71.7	83.2	72.4	83.7	-	-	84.7	91.5
DS-ComGAN	**60.7**	**71.3**	**74.5**	**84.6**	**76.7**	**86.6**	**76.9**	83.1	**90.0**	**94.6**

Table 3: **Segmentation results on training data measured in IoU and Dice**. DS-ComGAN is compared with the state-of-the-art un(weakly-)supervised segmentation methods. Following the [44], † indicates unfair baseline results obtained using extra ground-truth information. ⋇ represents a GAN-based model. OneGAN⋄ is a weakly supervised baseline, which requires clean backgrounds as additional inputs.

※ DS-ComGAN exhibits:

- **Unsupervised object segmentation;**
- **More precise segmentation**(IoU↑ and Dice↑);
- **Robustness to diverse datasets.**

Thank you !