ComGAN: Unsupervised Disentanglement and Segmentation via Image Composition

Rui Ding
Central South University
ruiding@csu.edu

Kehua Guo*
Central South University
guokehua@csu.edu

Xiangyuan Zhu
Central South University
zhuxiangyuan@csu.edu

Zheng Wu
Central South University
wuzhenghuse@gmail.com

Liwei Wang
Central South University
wang.liwei@csu.edu

Abstract

We propose ComGAN, a simple unsupervised generative model, which simultaneously generates realistic images and high semantic masks under an adversarial loss and a binary regularization. In this paper, we first investigate two kinds of trivial solutions in the compositional generation process, and demonstrate their source is vanishing gradients on the mask. Then, we solve trivial solutions from the perspective of architecture. Furthermore, we redesign two fully unsupervised modules based on ComGAN (DS-ComGAN), where the disentanglement module associates the foreground, background and mask with three independent variables, and the segmentation module learns object segmentation. Experimental results show that (i) ComGAN's network architecture effectively avoids trivial solutions without any supervised information and regularization; (ii) DS-ComGAN achieves remarkable results and outperforms existing semi-supervised and weakly supervised methods by a large margin in both the image disentanglement and unsupervised segmentation tasks. It implies that the redesign of ComGAN is a possible direction for future unsupervised work.[2]

1 Introduction

Generative adversarial networks (GANs) [1] have been successful in realistic image generation [2, 3], and recent work on GAN-based image composition [4, 5, 6, 7, 8, 9] has shown that GANs identify and disentangle a class of objects from the background. To direct that GANs learn the distinction between foreground and background, this kind of work commonly requires the following assumption:

Assumption 1 (Image composition [10]) *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:*

$$x = x_f \odot x_m + x_b \odot (1 - x_m), \tag{1}$$

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

It is notable that the Assumption 1 is mild and achieves significant outcomes in many areas. For example, the models [11, 12] capture meaningful foreground segmentation masks via exactly coupling different foregrounds and backgrounds. C3-GAN [13] proposes a scene decomposition-based method to enforce the model to learn the features of foreground objects. Yang et al. [14, 15] detect moving

*Corresponding author

[2]Code and data are available at https://github.com/Ruiding1/ComGAN

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

objects in video by minimizing the mutual information of foreground-background and training in an adversarial manner. A class of methods [16, 17] explores and perturbs the latent space of pre-trained GANs to find foreground masks. However, it is observed that under the Assumption 1, the generation process is always accompanied by trivial solutions [10, 18, 11, 19, 20, 21, 22, 12]. Trivial solutions [11] can be considered as meaningless masks generated by models. To our knowledge, existing works do not indicate that the source of the trivial solutions and they alleviate this issue in two ways. One way is to add supervised information [23, 21, 19, 18], such as CGN [19] avoids trivial solutions by adding pre-trained U2-Net [24]. Another is to design clever regularization and fine-tune the parameters [10, 11, 20, 22, 12]. For example, PerturbGAN [11] proposes a regularization based on image recombination with arbitrarily small relative shifts. Yang et al. [20] propose a regularization based on mutual information maximization. These methods are sensitive to object scale, object category and datasets, and therefore consume computational resources and time costs to perform hyperparameters fine-tuning. For our solution, we propose an unsupervised generative model called ComGAN, which simultaneously generates realistic images and high semantic masks. In addition, our model effectively avoids trivial solutions from the perspective of architecture, which implies that the model does not require any supervised information and explicit regularization. More specifically, ComGAN is a generic way to generalize two typical image compositional generation methods and alleviate the shortcomings coming from the above two methods. Furthermore, we place unique restrictions on each module, which help the model avoid trivial solutions.

Trivial solutions also have a negative role in related tasks. For example, in the image disentanglement, low semantic masks cause foreground and background conflicts [25, 26], which reduce the synthesis quality. In unsupervised segmentation, the regularization to alleviate trivial solutions may induce overfitting of the segmentation masks [14, 15, 12]. To highlight the flexibility and robustness of ComGAN, we extend it to the above two tasks. The ComGAN-based variant is called DS-ComGAN, which contains both disentanglement and segmentation modules. For multi-factor disentanglement, existing methods [26, 21, 27] rely on additional supervised information to learn the distinction of image regions. Notice that ComGAN achieves foreground-background disentanglement in an unsupervised way. As a result, the disentanglement module is designed as follows: we add more global information to the shared features in ComGAN and maximize the mutual information between variables and images. This module simplifies the previous hierarchical generative network and outperforms the state-of-the-art (SOTA) semi-supervised and weakly supervised image disentanglement methods. For unsupervised segmentation, existing methods [14, 15, 11, 12] rely on strong assumptions, such as that the foreground and background are largely independent, which limits their applicability. Different from these methods, we train a segmentation network using the images and semantic masks synthesized by the disentanglement module. By adversarial training strategy, the image distribution and the mask distribution are aligned. Furthermore, a consistency regularization is introduced to ensure that the predicted masks are consistent with the inputs. This segmentation module relies only on mild Assumption 1 and outperforms the SOTA unsupervised segmentation methods.

The main contributions are highlighted as follows:

1. To the best of our knowledge, we are the first to find that the source of trivial solutions is vanishing gradients on the mask. Furthermore, we propose ComGAN, a simple unsupervised generative model, which is the first to solve trivial solutions from the perspective of architecture and achieve foreground-background disentanglement with only an adversarial loss and a binary regularization.

2. We propose DS-ComGAN, a variant network based on ComGAN, which achieves image disentanglement and object segmentation in a fully unsupervised manner. Experiments show that DS-ComGAN is robust to various datasets and outperforms SOTAs in both tasks.

2 Related Work

GAN-based Image Composition. Image composition can be regarded as combining multiple visual areas to construct a realistic image [28]. A series of typical methods built on GANs [1] perform image composition by utilizing Assumption 1 to achieve various functions. LR-GAN [10] composites realistic images by learning to generate the backgrounds and foregrounds separately and recursively. FineGAN [25] hierarchically composites images, which disentangles the background, object shape, and object appearance. PerturbGAN [11] learns segmentation masks by training a generative model of a layered scene and composing foregrounds and backgrounds. Similarly, SEIGAN [18] performs

the cut, paste and inpaint operations consistent with semantic information to obtain segmentation masks. Although the compositional image generation performs excellently in several tasks under Assumption 1, these methods require additional regularization and hyperparameters tuning to avoid trivial solutions.

Controllable Image Disentanglement. The goal of controllable image disentanglement is to disentangle factors of variation (e.g., object shape and object appearance) towards controllable generation. InfoGAN [29] proposes semantic image generation control by imposing regularization of mutual information. A series of works [25, 30, 31, 26] hierarchically generates images and disentangles the background, the object's shape and its appearance by bounding box annotation. MixNMatch [31] extends the FineGAN [25], which disentangles and encodes four variable factors (background, object pose, shape, and texture) from real images. Along this direction, PartGAN [30] achieves part-level decomposition by learning a part generator. Benny and Wolf [21] further propose a complex GAN-based generative model, which requires a set of clean background images to simultaneously solve several tasks including image disentanglement and unsupervised segmentation. These methods [32, 33] learn the independent latent characteristics of an object, especially its appearance and pose. CGN [19] decomposes the image generation process into independent causal mechanisms, and allows for generating counterfactual images. SSC-GAN [27] is a semi-supervised single-stage generative model considers three factors of variation via class labels, namely independent variables, cross-class variables, and class variables. Although these methods exhibit strong performance, they all rely on additional supervised information to achieve multi-factor disentanglement.

Unsupervised Object Segmentation. Unsupervised object segmentation aims to extract a useful interpretation from an image in an unsupervised way. W-net [34] extracts foreground objects through an encoder-decoder framework and minimizes reconstruction errors. ReDO [35] captures the mask of the object, relying on the assumption that changing the foreground objects does not affect the overall data distribution. Assuming that the foreground and background are largely independent, these methods [14, 15, 11, 12] obtain segmentation masks by reconstructing the realistic images, i.e., precisely combining the different foregrounds and backgrounds. Labels4Free [12] performs unsupervised object segmentation based on the idea that exact coupling between foreground and background is highly nontrivial. A similar assumption is implemented In method [36], where IEM learns image segmentation from an information information-theoretic by maximizing inpainting error. Similarly, Yang et al. [20] synthesize paired realistic images and segmentation masks by maximizing mutual information between generated images and latent variables. These methods frequently require stringent assumptions and are not necessarily applicable to complex real-world data.

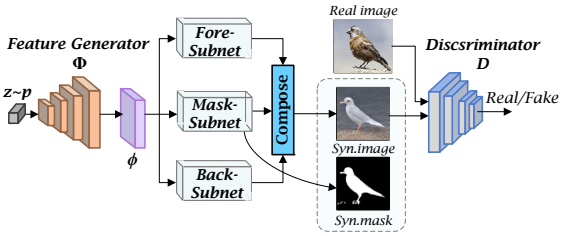
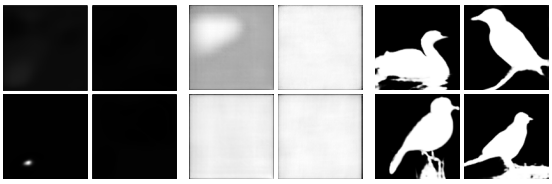


Figure 1: **Overview of ComGAN.** The generator of ComGAN is composed of the feature generator Φ and three independent subnetworks. Three independent subnetworks synthesize the foreground, background and mask with features ϕ as a shared input. The 'compose' indicates the utilization of Assumption 1 to compose the image. The discriminator D distinguishes real images from fake ones.



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

Figure 2: **Comparison results of the synthesized mask.** ComGAN synthesizes high semantic masks, while the masks synthesized by models Π_1 and Π_2 both converge to trivial solutions. Note that Π_1 has the same structure as [25, 30, 31, 26], and Π_2 has a similar structure to [11, 19]. All models include identical adversarial loss and binary regularization βL_{binary} from [19]. For each model: (left) β = 0.1, (right) β= 1.

3 Proposed Approach

3.1 Trivial solutions and ComGAN

Source of trivial solutions. Two kinds of trivial solutions were introduced in [11]. The first trivial solution degrades the whole scene into foreground or background, and the mask is always full or empty. The second trivial solution makes the foreground and background identical, and the mask is meaningless for the composite scene. Then, we state the source of two kinds of trivial solutions.

Lemma 1 ***(Vanishing gradients on the mask)****. Let L_{all} be the overall loss function and $\bar{x}_m$ be the synthesized mask. Consider a model that composes an image $\bar{x}$ utilizing Assumption 1. There exist vanishing gradients on the mask, i.e., $\partial L_{all}/\partial \bar{x}_m = 0$ if and only if the model converges to two kinds of trivial solutions.*

Proof: See Appendix B.1. □

Network architecture of ComGAN. An overview of our proposed ComGAN is illustrated in Fig. 1. The generator of ComGAN consists of two parts. The feature generator denoted as Φ synthesizes features conditioned on the variable z, i.e., $\phi = \Phi(z)$ where z is sampled from a prior distribution p_0. The three independent subnetworks are defined as $\mathcal{F}$, $\mathcal{B}$ and $\mathcal{M}$. They synthesize the foreground, background and mask, that is, $\bar{x}_f = \mathcal{F}(\phi)$, $\bar{x}_b = \mathcal{B}(\phi)$ and $\bar{x}_m = \mathcal{M}(\phi)$. Driven by Assumption 1, the composite image $\bar{x}$ is defined as:

$$\bar{x} = \mathcal{F}(\Phi(z)) \odot \mathcal{M}(\Phi(z)) + \mathcal{B}(\Phi(z)) \odot (1 - \mathcal{M}(\Phi(z))) \tag{2}$$

Remark 1 *This form generalizes two typical image compositional generation methods:*

- *If only $\Phi(\cdot)$ in $\mathcal{B}$ is an identity map, i.e., $\mathcal{B}(\Phi(z)) = \mathcal{B}(z)$, then this form is equivalent to model Π_1, that is, two independent generators synthesize a composite image where shared features exist in the foreground and mask generation.*
- *If $\Phi(\cdot)$ is an identity map, i.e., $\Phi(z) = z$, then this form is equivalent to model Π_2, that is, three independent generators synthesize a composite image where foreground, background and mask are generated by three generators respectively.*

From the above observations, we notice that the two typical methods have a common shortcoming, that is, models Π_1 and Π_2 both contain an independent background generation process. This term ignores the association of foreground, background and mask on the features. Moreover, if the model converges to trivial solutions, it implies that the model degrades to an original GAN that suffers from known shortcomings, such as mode collapse. As a result, our proposed method is a generic way and mitigates the above shortcoming. Then, the adversarial training loss L_D^{adv} is defined as follows:

$$\min_{\Phi,\mathcal{F},\mathcal{B},\mathcal{M}} \max_{D} \mathcal{L}_D^{adv} = \mathbb{E}_{x\sim p_{\text{data}}}\left[\log\left(D(x)\right)\right] + \mathbb{E}_{z\sim p_0}\left[1 - \log D(\bar{x})\right]. \tag{3}$$

Obviously, Lemma 1 points us to the limitation of the compositional generation under Assumption 1. Then, we illustrate how to address the issue from the perspective of architecture.

Theorem 1 *Given a generation G_θ composed of a decoder Φ_{θ_ϕ}: $\mathcal{Z} \to \phi$ and three subnets $\mathcal{F}_{\theta_f}$, $\mathcal{B}_{\theta_b}$ and $\mathcal{M}_{\theta_m}$: $\phi \to \mathcal{X}$. Let D be a discriminator and $D^*(G^*(\cdot))$ be Nash equilibrium. If the composite images satisfy $\bar{x} = \mathcal{F}(\Phi(z)) \odot \mathcal{M}(\Phi(z)) + \mathcal{B}(\Phi(z)) \odot (1 - \mathcal{M}(\Phi(z)))$, $\|D(G(\cdot)) - D^*(G^*(\cdot))\| < \epsilon$, $\max\{\mathbb{E}_{\phi\sim p(\Phi(z))}\left[\left\|J_{\theta_f}\mathcal{F}(\phi)\right\|_2^2\right], \mathbb{E}_{\phi\sim p(\Phi(z))}\left[\left\|J_{\theta_b}\mathcal{B}(\phi)\right\|_2^2\right]\} \leq \delta^2$ and $\|\nabla_\theta L_D^{adv}\|_2 \geq \sigma$, then*

$$\left\|\nabla_{(\theta_\phi,\theta_m)}\mathbb{E}_{z\sim p(z)}\left[\log\left(1 - D\left(\mathcal{F}(\Phi(z))\right)\right)\right]\right\|_2^2 > \sigma^2 - \frac{\delta^2\epsilon^2}{(1/2-\epsilon)^2}. \tag{4}$$

Proof: See Appendix B.1. □

We denote that $\rho = \sigma^2 - \frac{\delta^2\epsilon^2}{(1/2-\epsilon)^2}$ and trivial masks as $\bar{x}_m^*$. If $\rho > 0$, then the Φ and $\mathcal{M}$ are updated. The updated masks are $\bar{x}_m^+ = \mathcal{M}_{\theta_m^+}(\Phi_{\theta_\phi^+}(z))$, which means that the model can escape from the first trivial solution, i.e. $\bar{x}_m^+ \neq \bar{x}_m^*$ and $\partial L_D^{adv}/\partial \bar{x}_m^+ \neq 0$.

Corollary 1 *The following modules restrictions help the model avoid trivial solutions: the $\mathcal{F}$ and $\mathcal{B}$ are lightweight and differential, the $\mathcal{M}$ is a shallow network and the capacity of Φ is enough.*

Proof: See Appendix B.1. □

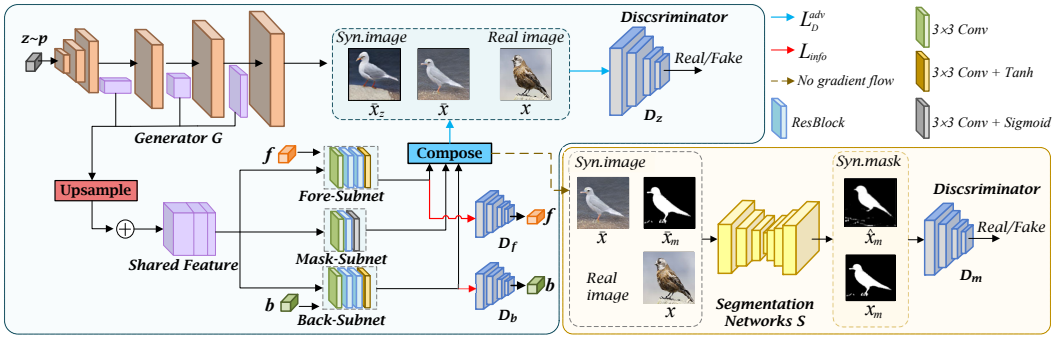


Figure 3: **Overview of DS-ComGAN.** DS-ComGAN consists of two modules: (i) the disentanglement module (blue); (ii) the segmentation model (yellow). The generator G synthesizes $\bar{x}_z$ to perform adversarial training with the discriminator D_z to contain more global features. Shared features are extracted from G and fed to three independent subnetworks. Similar to ComGAN, foreground, background, and mask are synthesized and composited into an image $\bar{x}$ based on the Assumption 1. The module implements image disentanglement by the adversarial loss L_D^{adv} and the mutual information regularization L_{info}. Since our model synthesize high semantic masks, the segmentation module performs unsupervised segmentation with image distribution alignment and mask distribution alignment via adversarial training.

3.2 The Disentanglement Module of DS-ComGAN

Image generation is controlled by the mask variable z, foreground variable f and background variable b, where f and b are both N dimensional random one-hot vectors. The generator of This module consists of two parts, namely a feature generator denoted as G and three independent subnetworks defined as $\mathcal{F}$, $\mathcal{B}$ and $\mathcal{M}$.

Global information for shared features. The features should contain global information so that the model captures a wide range of distinctions between foreground and background, rather than focusing on abrupt regions. To add global information to the shared features of G, we synthesize realistic images for adversarial training, i.e., $\bar{x}_z$ is synthesized by $G(z)$ and fed to the discriminator D_z. Then, we extract the activation map from the last three layers inside G, denoted as a_1, a_2 and a_3. The shared features with global information can be written as follows:

$$\phi = \mathbb{U}(a_1) \oplus \mathbb{U}(a_2) \oplus \mathbb{U}(a_3), \tag{5}$$

where $\mathbb{U}$ is a spatially upsample to maintain the same size for all activation maps and $\oplus$ is a concatenation along the channel dimension. In order to control image disentanglement, $\mathcal{F}$, $\mathcal{B}$ and $\mathcal{M}$ synthesize foreground, background, and mask conditioned on f, b and z, i.e., $\bar{x}_f = \mathcal{F}(\phi, f)$, $\bar{x}_b = \mathcal{B}(\phi, b)$ and $\bar{x}_m = \mathcal{M}(\phi)$. Driven by Assumption 1, the composite image $\bar{x}$ are defined as:

$$\bar{x} = \mathcal{F}(\phi, f) \odot \mathcal{M}(\phi) + \mathcal{B}(\phi, b) \odot (1 - \mathcal{M}(\phi)). \tag{6}$$

Then, the discriminator D_z is trained to classify images as real or fake and the adversarial training loss $L_{D_z}^{adv}$ is defined as follows:

$$\min_{G,\mathcal{F},\mathcal{B},\mathcal{M}} \max_{D_z} \mathcal{L}_{D_z}^{adv} = \mathbb{E}_x\left[\log D_z(x)\right] + \mathbb{E}_z\left[\log(1 - D_z(\bar{x}_z))\right] + \mathbb{E}_{z,f,b}\left[\log(1 - D_z(\bar{x}))\right]. \tag{7}$$

Controllable regularization for foreground-background subnetworks. A mutual information regularization is incorporated into the $\mathcal{F}$ and $\mathcal{B}$ to associate the variables f and b with the synthesized $\bar{x}_f$ and $\bar{x}_b$. Similar to InfoGAN [29], we obtain similar results by maximizing the lower bound on the variance of the mutual information.

$$\max_{D_f, D_b} L_{info} = I(\bar{x}_f, f) + I(\bar{x}_b, b) \geq \mathbb{E}_{z,f,b}[\log D_f(f \mid \bar{x}_f) + \log D_b(b \mid \bar{x}_b)], \tag{8}$$

where $I(\cdot)$ is the mutual information, two discriminators D_f and D_b approximate posterior distribution of $p(f|\bar{x}_f)$ and $p(b|\bar{x}_b)$. As implemented in [37, 20, 25], the $I(\cdot)$ is regarded as a regularization of a class of unsupervised clusters. In other words, D_f guides $\mathcal{F}$ to focus on generating foreground regions and learning the features of objects. The similar analysis applies to D_b. According to [10, 38], it is difficult for GANs to learn a set of disconnected manifolds. Intuitively, $\mathcal{F}$ and $\mathcal{B}$ learn simpler and continuous distributions, and thereby L_{info} improves the quality of image generation.

3.3 The Segmentation Module of DS-ComGAN

The following two networks are defined: a segmentation network $\mathcal{S}$ predicts segmentation masks conditioned on the input images; a discriminator D_m classifies masks as real or fake. Our key idea is to exploit the high semantic masks synthesized by disentanglement module to align the image distribution and the mask distribution via adversarial training.

Image distribution alignment. In the disentanglement module, D_z is trained to identify x sampled from p_{data} and $\bar{x}$ composited by our model. When the model achieve Nash equilibrium, $\bar{x}$ has the identical distribution as x, i.e., $p_{data}(x) = p(\bar{x})$. Hence, image distribution alignment is accomplished.

Mask distribution alignment. In the first step, we input the real image x and the synthetic image $\bar{x}$ into the segmentation network to obtain the corresponding prediction masks, i.e., $x_m = \mathcal{S}(x)$ and $\hat{x}_m = \mathcal{S}(\bar{x})$. Then, a mask triplet ($\bar{x}$, $\hat{x}$, x_m) is fed into the discriminator D_m where the synthesized mask x_m is regarded as real and both the predicted masks $\hat{x}$ and x_m are regarded as fake. Mask distributions of x_m, $\hat{x}_m$ and $\bar{x}_m$ are aligned via adversarial training,

$$\min_{\mathcal{S}} \max_{D_m} \mathcal{L}_{D_m}^{adv} = \mathbb{E}_{\bar{x}_m, \hat{x}_m, x_m}[\log D_m(\bar{x}_m) + \log(1 - D_m(\hat{x}_m)) + \log(1 - D_m(x_m))]. \quad (9)$$

The mask distribution alignment makes the predicted masks clearer, but this adversarial training strategy causes the issue that predicted masks are inconsistent with the input images. Hence, the corresponding consistency loss L_{cons} is introduced to ensure that the predicted masks are consistent with the inputs, rather than merely learning the mask distribution.

$$\min_{\mathcal{S}} \mathcal{L}_{cons} = \mathbb{E}_{\bar{x}, \bar{x}_m} \left[\| \bar{x}_m - \mathcal{S}(\bar{x}) \| \right]. \quad (10)$$

We transform the unsupervised segmentation task into two adversarial tasks and acquire segmentation masks by enhancing image and mask distribution consistency.

3.4 Model Training

The whole training process of ComGAN is simple, with only an adversarial loss and a binary regularization. The learning objective is as follows:

$$L_{all} = \min_{\Phi, \mathcal{F}, \mathcal{B}, \mathcal{M}} \max_{D} \mathcal{L}_{D}^{adv} + \min_{\Phi, \mathcal{M}} \beta \mathcal{L}_{\text{binary}}, \quad (11)$$

where $\mathcal{L}_{\text{binary}}$ is the pixel-wise binary entropy of mask from [19]. An explicit formulation of $\mathcal{L}_{\text{binary}}$ is shown in Appendix B.2. The sharpness of the semantic mask is modified by fine-tuning β. The results of the visualization on fine-tuning β are available in Appendix C.1. All the constituent networks of DS-ComGAN perform two kinds of unsupervised tasks: image disentanglement and object segmentation. Overall, we optimize the following loss:

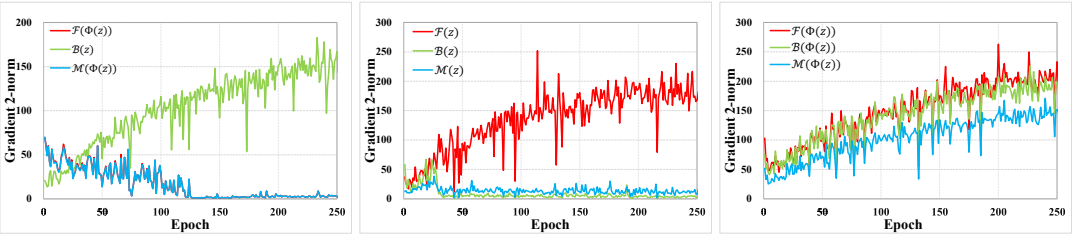
$$L_{all} = \overbrace{\max_{D_f, D_b} L_{info} + \underbrace{\min_{G, \mathcal{F}, \mathcal{B}, \mathcal{M}} \max_{D_z} \mathcal{L}_{D_z}^{adv} + \min_{\mathcal{S}} \max_{D_m} \mathcal{L}_{D_m}^{adv} + \min_{\mathcal{S}} \lambda \mathcal{L}_{\text{cons}}}_{\text{objective for object segmentation task}}}^{\text{objective for image disentanglement task}}. \quad (12)$$

For the image disentanglement task, the model is performed in one stage. Then, the synthesized images and semantic masks are fed into the segmentation network. Hence, the objective segmentation process is composed of two stages.

4 Experiments

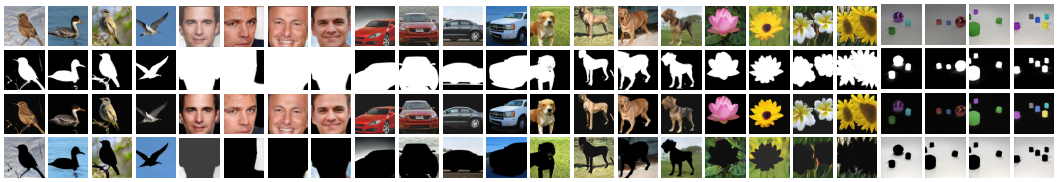
4.1 Experimental Setup

Datasets and implementation details. The experiments are conducted on five fine-grained image datasets and a multi-object dataset: CUB [39], FS-100 [40], Stanford-Cars [41]. Stanford-Dogs [41], Flowers [42], CLEVR6 [43]. The ground truth segmentation masks of Stanford-Cars and Stanford-Dogs are approximated by following the practice in [21, 44, 20]. More details about datasets and the implementation are available in the supplementary material.



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

Figure 4: **Comparison results of gradient 2-norm during the training.**



(a) CUB (b) FS100* (c) Stanford-Cars (d) Stanford-Dogs (e) Flowers (f) CLEVR6

Figure 5: **Qualitative generation results for each dataset**. Zoom in for better visibility. From top to bottom: (i) final images; (ii) masks; (iii) foregrounds; (iv) backgrounds. * indicates that we incorporate a binary regularization to obtain a high semantic mask, since the foreground almost covers the entire image.

Baseline and evaluation protocol. In order to analyze the effectiveness of the adopted components, we established a baseline model, i.e. SimpleGAN [25]. SimpleGAN aims to generate a realistic image, and its network contains a generator and a discriminator. For a fair comparison, we use SimpleGAN as the feature generator of ComGAN and DS-ComGAN. We assess synthesis quality in terms of Inception Score (**IS**) [45] and Fréchet Inception Distance (**FID**) [46], which are computed on 20K randomly synthesized images. To quantitatively evaluate the quality of the predicted masks, both the Intersecion of Union (**IoU**) and Dice score (**Dice**) [47] are used as the evaluation metrics.

4.2 Trivial solutions avoidance based on ComGAN

Avoidance of trivial solutions. To show that ComGAN solves trivial solutions, we select two typical networks represent model Π_1 and Π_2 for comparison. All models only include identical adversarial loss and binary regularization. As a typical network for model Π_1, FineGAN has two generator where a generates foreground and mask and another generates background. PerturbGAN is selected as the typical network for model Π_2, which consists of three generators that generates foreground, background and mask, respectively. For a fair comparison, we make the following two things: (i) we use the code provided by the authors in FineGAN and retrain it with the identical training parameters; (ii) we simplify StyleGAN [48] with SimpleGAN in PerturbGAN and retrain it. As shown in Fig. 2, the mask synthesized by ComGAN avoids convergence to trivial solutions and has non-trivial semantic information. In addition, we track the gradient 2-norm of the three models. It is clear from Fig. 4 that the gradient norms of mask networks in model Π_1 and model Π_2 converge to zero and the model degrades to an original GAN, while our method effectively avoids vanishing gradients.

4.3 Image Disentanglement based on DS-ComGAN

Quantitative evaluation of image generation. We compare DS-ComGAN with the competing GAN-based generative models without any advanced GAN training strategy. As illustrated in Table 1, the results show that DS-ComGAN significantly outperforms all the SOTA weakly supervised and semi-supervised methods. FineGAN and MixNMatch may be limited by the foreground-background conflict, such as trivial masks. SSC-GAN might have trouble in learning a set of disconnected manifolds. To emphasize that DS-ComGAN is fully unsupervised and adapts to complex multi-object scenarios, we further perform our model on diverse datasets. Stanford-Dogs and Flowers typically include a single foreground object, whereas CLEVR6 contains a variety of occluded and truncated objects. Table 3 shows that our model is robust on various datasets and outperforms the SOTA image

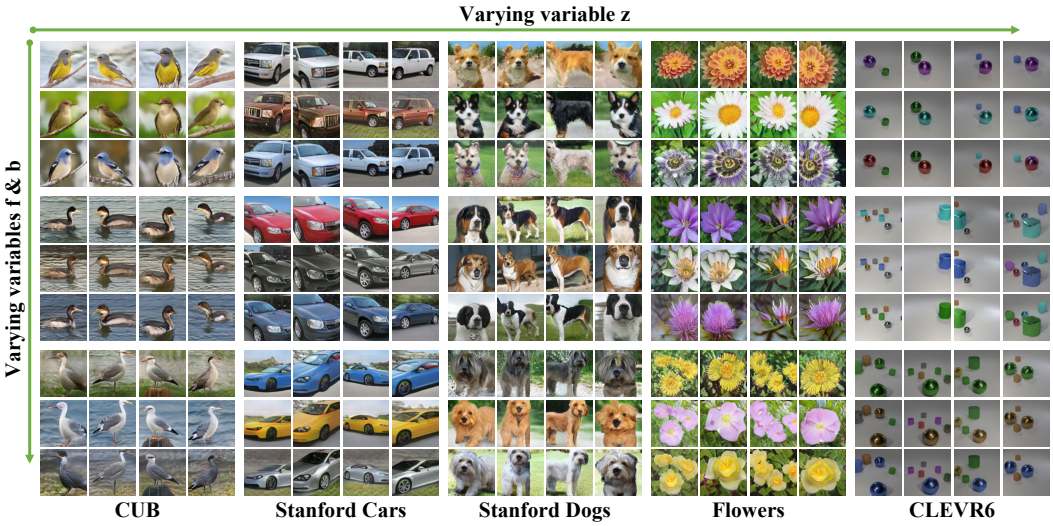


Figure 6: **The controllable image generation via varying variables f, b and z.** Zoom in for better visibility.

Methods	Sup.	CUB		FS-100		Stanford-Cars	
		FID ↓	IS↑	FID ↓	IS↑	FID ↓	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.

disentanglement methods. Fig 5 shows the qualitative generation results for each dataset, from which it is evident that our model generates not only realistic images but also clear semantic masks.

Controllable image disentanglement. Fig. 6 illustrates that by varying three independent variables, DS-ComGAN synthesizes realistic and diverse images. This indicates that the model achieves controlled image disentanglement. We see that variable z is associated with the masks. The pose of the object is altered with the varying variable z. The colors and textures of the foreground and background are associated with variables f and b. To clarify the role of f and b in image disentanglement, Fig 7 further analyzes the disentanglement of foreground and background. In summary, DS-ComGAN associates three independent variables (f, b and z) with the foreground, background and mask, respectively.

(a) Varying b, fixed f and z (b) Varying f, fixed b and z

Figure 7: **Disentanglement of foreground-background.**

Method	FID ↓	IS ↑
Baseline	89.87	4.45±0.46
+ Independent subnetworks.	21.35	4.75±0.08
+ L_{info}-based Reg.	17.32	4.73±0.08
+ Extract global features.	16.26	4.79±0.47
Improvement	**73.61**	**0.34**

Table 2: **The results of the baseline and variants on CUB.**

	Single Object				Multi-Object	
	Stanfor-Dogs		Flowers		CLEVR6	
Methods	FID ↓	IS↑	FID ↓	IS↑	FID ↓	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 3: **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.

(a) CUB (b) Stanford-Dogs (c) Stanford-Cars (d) Flowers (e) CLEVR6

Figure 8: **Qualitative segmentation results for each dataset.** From left to right: (i) observed images, (ii) ground-truth masks, (iii) predicted masks. More samples can be found in the supplementary material.

Effectiveness of model components To evaluate the contribution of each proposed component, we analyze quantitatively the performance margin between the baseline model and ComGAN. As demonstrated in Table 2, the incorporation of $\mathcal{F}$, $\mathcal{B}$ and $\mathcal{M}$ significantly improves the performance of the baseline model. It illustrates that ComGAN has an excellent image generation performance. As analyzed in section 3.2, the L_{info} causes a decrease in FID by about 3 points.

4.4 Unsupervised Segmentation based on DS-ComGAN

To demonstrate that our method applies to complex real-world domains and multi-object scenarios with challenging spatial configurations, along the lines of [44], DS-ComGAN performs the unsupervised segmentation task on the various datasets. Then, we perform a comparison between DS-ComGAN and several state-of-the-art un(weakly-)supervised segmentation methods.

Qualitative analysis. In Fig. 8, qualitative segmentation results of DS-ComGAN are presented. We observe that the predicted masks segment visual details more precisely than the ground-truth masks, such as the legs of birds and the rearview mirrors of cars. Although the predicted masks properly detect and segment foreground objects on the Flowers and Stanford-Cars datasets, the masks are inconsistent with the foreground images, which is mitigated by increasing λ.

Quantitative comparisons to prior work. Table 4 demonstrates that DS-ComGAN exhibit robustness to diverse datasets and achieve superior performance across all metrics and datasets. This implies that the model pays more attention to the discrepancy between the objects and background rather than the number of objects. In addition, the result benefits from the images and high semantic masks synthesized by DS-ComGAN, which ensure two high-quality distribution alignments.

	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 4: **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.

5 Conclusion

As the limitation of method, DS-ComGAN has struggled to achieve the desired performance when the foreground object features are highly diverse (e.g., HKU-IS [56]). We conjecture that the limitation might relate to the capacity of model. DS-ComGAN performs excellently in controlled image synthesis tasks, which may cause the incidence of image falsification.

We first analyze that the source of trivial solutions is vanishing gradients on the mask. Subsequently, we propose ComGAN, a simple unsupervised generative model, which simultaneously generates realistic images and high semantic masks, and effectively avoids trivial solutions. Then, we design DS-ComGAN based on ComGAN, which exhibits excellent performance in both image disentanglement and unsupervised segmentation tasks. Notably, DS-ComGAN obtains these excellent performances by relying on its architecture rather than complex and clever regularization. It implies that ComGAN has the potential to shine in many tasks.

Acknowledgments

This work was supported by the Natural Science Foundation of China under Grant 62076255 and 62177047; Open Research Projects of Zhejiang Lab (NO. 2022RC0AB07); Hunan Provincial Science and Technology Plan Project 2020SK2059; Key projects of Hunan Education Department 20A88; National Science Foundation of Hunan Province 2021JJ30082, 2019JJ20025 and 2019JJ40406; National Social Science Fund of China (No. 20&ZD120).

References

[1] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[2] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[3] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[4] Klaus Greff, Antti Rasmus, Mathias Berglund, Tele Hao, Harri Valpola, and Jürgen Schmidhuber. Tagger: Deep unsupervised perceptual grouping. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[5] SM Eslami, Nicolas Heess, Theophane Weber, Yuval Tassa, David Szepesvari, Geoffrey E Hinton, et al. Attend, infer, repeat: Fast scene understanding with generative models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[6] Bor-Chun Chen and Andrew Kae. Toward realistic image compositing with adversarial learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[7] Vladimir V Kniaz, Vladimir Knyaz, and Fabio Remondino. The point where reality meets fantasy: Mixed adversarial generators for image splice detection. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[8] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[9] Chen-Hsuan Lin, Ersin Yumer, Oliver Wang, Eli Shechtman, and Simon Lucey. St-gan: Spatial transformer generative adversarial networks for image compositing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[10] Jianwei Yang, Anitha Kannan, Dhruv Batra, and Devi Parikh. Lr-gan: Layered recursive generative adversarial networks for image generation. In *International Conference on Learning Representations (ICLR)*, 2017.

[11] Adam Bielski and Paolo Favaro. Emergence of object segmentation in perturbed generative models. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[12] Rameen Abdal, Peihao Zhu, Niloy J Mitra, and Peter Wonka. Labels4free: Unsupervised segmentation using stylegan. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[13] Yunji Kim and Jung-Woo Ha. Contrastive fine-grained class clustering via generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2022.

[14] Yanchao Yang, Antonio Loquercio, Davide Scaramuzza, and Stefano Soatto. Unsupervised moving object detection via contextual information separation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[15] Yanchao Yang, Brian Lai, and Stefano Soatto. Dystab: Unsupervised object segmentation via dynamic-static bootstrapping. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[16] Andrey Voynov and Artem Babenko. Unsupervised discovery of interpretable directions in the gan latent space. In *Proceedings of International Conference on Machine Learning (ICML)*, 2020.

[17] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Finding an unsupervised image segmenter in each of your deep generative models. *arXiv preprint arXiv:2105.08127*, 2021.

[18] Pavel Ostyakov, Roman Suvorov, Elizaveta Logacheva, Oleg Khomenko, and Sergey I Nikolenko. Seigan: Towards compositional image generation by simultaneously learning to segment, enhance, and inpaint. *arXiv preprint arXiv:1811.07630*, 2018.

[19] Axel Sauer and Andreas Geiger. Counterfactual generative networks. In *International Conference on Learning Representations (ICLR)*, 2021.

[20] Learning foreground-background segmentation from improved layered gans. In *Proceedings of the IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2022.

[21] Yaniv Benny and Lior Wolf. Onegan: Simultaneous unsupervised learning of conditional image generation, foreground segmentation, and fine-grained clustering. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2020.

[22] Sébastien Ehrhardt, Oliver Groth, Aron Monszpart, Martin Engelcke, Ingmar Posner, Niloy Mitra, and Andrea Vedaldi. Relate: Physically plausible multi-object scene synthesis using structured latent spaces. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[23] Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[24] Xuebin Qin, Zichen Zhang, Chenyang Huang, Masood Dehghan, Osmar R Zaiane, and Martin Jagersand. U2-net: Going deeper with nested u-structure for salient object detection. *Pattern Recognition*, 106:107404, 2020.

[25] Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[26] Utkarsh Ojha, Krishna Kumar Singh, and Yong Jae Lee. Generating furry cars: Disentangling object shape & appearance across multiple domains. In *International Conference on Learning Representations (ICLR)*, 2021.

[27] Tianyi Chen, Yi Liu, Yunfei Zhang, Si Wu, Yong Xu, Feng Liangbing, and Hau San Wong. Semi-supervised single-stage controllable gans for conditional fine-grained image generation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2021.

[28] Li Niu, Wenyan Cong, Liu Liu, Yan Hong, Bo Zhang, Jing Liang, and Liqing Zhang. Making images real again: A comprehensive survey on deep image composition. *arXiv preprint arXiv:2106.14490*, 2021.

[29] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2016.

[30] Yuheng Li, Krishna Kumar Singh, Yang Xue, and Yong Jae Lee. Partgan: Weakly-supervised part decomposition for image generation and segmentation. In *Proceedings of British Machine Vision Conference (BMVC)*, 2021.

[31] Yuheng Li, Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Mixnmatch: Multifactor disentanglement and encoding for conditional image generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[32] Zhixin Shu, Mihir Sahasrabudhe, Riza Alp Guler, Dimitris Samaras, Nikos Paragios, and Iasonas Kokkinos. Deforming autoencoders: Unsupervised disentangling of shape and appearance. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2018.

[33] Patrick Esser, Johannes Haux, and Bjorn Ommer. Unsupervised robust disentangling of latent characteristics for image synthesis. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2019.

[34] Xide Xia and Brian Kulis. W-net: A deep model for fully unsupervised image segmentation. *arXiv preprint arXiv:1711.08506*, 2017.

[35] Mickaël Chen, Thierry Artières, and Ludovic Denoyer. Unsupervised object segmentation by redrawing. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[36] Pedro Savarese, Sunnie SY Kim, Michael Maire, Greg Shakhnarovich, and David McAllester. Information-theoretic segmentation by inpainting error maximization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[37] Zinan Lin, Kiran Thekumparampil, Giulia Fanti, and Sewoong Oh. Infogan-cr and modelcentrality: Self-supervised model training and selection for disentangling gans. In *Proceedings of International Conference on Machine Learning (ICML)*, 2020.

[38] Mohammadreza Armandpour, Ali Sadeghian, Chunyuan Li, and Mingyuan Zhou. Partition-guided gans. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[39] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The Caltech-UCSD Birds-200-2011 Dataset, 2011.

[40] Si Wu, Guangchang Deng, Jichang Li, Rui Li, Zhiwen Yu, and Hau-San Wong. Enhancing triplegan for semi-supervised conditional instance synthesis and classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[41] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshops*, 2013.

[42] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Proceedings of Indian Conference on Computer Vision, Graphics & Image Processing (ICVGIP)*, 2008.

[43] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[44] Peiyu Yu, Sirui Xie, Xiaojian Ma, Yixin Zhu, Ying Nian Wu, and Song-Chun Zhu. Unsupervised foreground extraction via deep region competition. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[45] Hong-Wei Ng and Stefan Winkler. A data-driven approach to cleaning large face datasets. In *Proceedings of international conference on image processing (ICIP)*, 2014.

[46] 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. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2017.

[47] Lee Raymond Dice. Measures of the amount of ecologic association between species. *Ecology*, 26:297–302, 1945.

[48] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[49] Chongxuan Li, Taufik Xu, Jun Zhu, and Bo Zhang. Triple generative adversarial nets. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[50] Zhe Gan, Liqun Chen, Weiyao Wang, Yuchen Pu, Yizhe Zhang, Hao Liu, Chunyuan Li, and Lawrence Carin. Triangle generative adversarial networks. In *Proceedings of Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[51] Yi Liu, Guangchang Deng, Xiangping Zeng, Si Wu, Zhiwen Yu, and Hau-San Wong. Regularizing discriminative capability of cgans for semi-supervised generative learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[52] Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. In *International Conference on Learning Representations (ICLR)*, 2018.

[53] Carsten Rother, Vladimir Kolmogorov, and Andrew Blake. " grabcut" interactive foreground extraction using iterated graph cuts. *ACM transactions on graphics (TOG)*, 23:309–314, 2004.

[54] Klaus Greff, Raphaël Lopez Kaufman, Rishabh Kabra, Nick Watters, Christopher Burgess, Daniel Zoran, Loic Matthey, Matthew Botvinick, and Alexander Lerchner. Multi-object representation learning with iterative variational inference. In *Proceedings of International Conference on Machine Learning (ICML)*, 2019.

[55] Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Proceedings of Neural Information Processing Systems (NeurIPS)*, 2020.

[56] Guanbin Li and Yizhou Yu. Visual saliency detection based on multiscale deep cnn features. *IEEE Transactions on Image Processing (TIP)*, 25:5012–5024, 2016.

[57] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017.

[58] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Proceedings of European Conference on Computer Vision (ECCV)*, 2014.

[59] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, 2016.

[60] Francisco Massa Wan-Yen Lo Yuxin Wu, Alexander Kirillov and Ross Girshick. Detectron2. https://github.com/facebookresearch/detectron2.

[61] Maria-Elena Nilsback and Andrew Zisserman. Delving into the whorl of flower segmentation. In *Proceedings of British Machine Vision Conference (BMVC)*, 2007.

[62] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] See Section 3.1, Section 4.2, Section 4.3, Section 4.4.
   (b) Did you describe the limitations of your work? [Yes] See Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Section 1.
   (b) Did you include complete proofs of all theoretical results? [Yes] See the Section 3.1 and supplemental material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code and data are available at `https://github.com/Ruiding1/ComGAN`. We provide detailed instructions of how to reproduce the main results in the supplemental material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 4.1, Section 4.2, Section 4.3, Section 4.4 and supplemental material.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See supplemental material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Details on Datasets and Hyperparameters

A.1 Dataset Details

Caltech-UCSD Birds-200-2011. This dataset contains about 12K bird images of 200 classes annotated with bounding boxes. The ground truth segmentation masks of the foreground object are annotated by humans and provided in the official release [39]. As in the previous work [25, 31, 44], We use the provided bounding box to extract a center square from the image, and scale it to 128×128 pixels.

Stanford Dogs. This dataset consists of about 20K dog images of 120 classes annotated with bounding boxes [41]. Similar to [25, 31, 44], We use the provided bounding box to extract a center square from the image, and scale it to 128×128 pixels. Since the ground-truth segmentation masks are not provided, we adopt the approximate ground-truth masks, generated by Yu et al. [44]. The specific generation process is as follows. Yu et al. follow the practice in Benny and Wolf[21] to approximate ground-truth masks with Mask R-CNN[57], pre-trained on the COCO[58] dataset with a ResNet-101[59] backend. The pre-trained model is acquired from the detectron2 [60] toolkit. Along the filtering strategy of [44], there exist 5,024 images with a clear foreground-background setup and high-quality mask.

Stanford Cars. This dataset consists of about 16K car images of 196 classes annotated with bounding boxes [41]. We use the provided bounding box to extract a center square from the image, and scale it to 128×128 pixels. To approximate the ground-truth masks, we adopt a similar process and filtering strategy to the Stanford Dogs dataset. Finally, there exist 12,322 images for experimentation.

FS-100. This dataset consists of about 15K human face images from the 100 largest classes [40, 27] of FaceScrub[45]. FaceScrub comprises a total of 107,818 face images of 530 celebrities. For a fair comparison, we follow the [27] by scaling all the images to 64×64 pixels.

Flowers. This dataset consists of about 8K flower images of 102 classes [42]. The ground truth segmentation masks are created by an automated approach designed exclusively for segmenting flowers in color photographs [61]. We scale all the images to 128×128 pixels.

CLEVR6 This dataset is a subset of the CLEVR dataset [43] with masks generated by Greff et al. [54]. We take the first 8K samples from CLEVR and scale these samples to 128×128 pixels.

A.2 Hyperparameters and Training Details.

We adopt the ADAM optimizer[62] with a learning rate of $\varsigma = 2 \times 10^{-4}$ and momentum parameters of $\beta_1 = 0.5$ and $\beta_2 = 0.999$. Following FineGAN[25] and DRC[44], we augment the real image with a horizontal flip and a random resizing crop. The variable dimensions are set as follows: (1) CUB: $N = 200$; (2) FS-100: $N = 100$; (3) Stanford-Dogs: $N = 150$; (4) Stanford-Cars: $N = 200$; (5) Flowers: $N = 150$; (6) CLEVR6: $N = 20$. We empirically set $\lambda = 5$ defined in equation (12) for all datasets. The maximum training epochs are set to 500, where the disentanglement module is first trained for 450 epochs with a batch size of 32, and the segmentation module is second trained for 50 epochs with a batch size of 64. All experiments are performed under the environments of NVIDIA GeForce RTX 3090 GPUs and an Intel Xeon Gold 6130 CPU.

B Details on Proofs and Regularization

B.1 Proofs of things

Proof of Lemma 1: (Sufficiency) Under Assumption 1, the $\bar{x}$ at each pixel $i, j \in \mathbb{R}$ can be written as follows:

$$\bar{x}[i,j] = \bar{x}_f[i,j] \cdot \bar{x}_m[i,j] + \bar{x}_b[i,j] \cdot (1 - \bar{x}_m[i,j]), \tag{13}$$

where $\bar{x}_f$ is the foreground and $\bar{x}_b$ is the background. As the model converges to the first trivial solution, the composite image degenerates to the foreground or background, i.e. $\bar{x}[i,j] = \bar{x}_f[i,j]$ or $\bar{x}[i,j] = \bar{x}_b[i,j]$. It implies that the compositional generation process degenerates into a single foreground or background generation, which causes vanishing gradients on the mask.

Let $\bar{G}_{\theta_m}$ be masks generator and θ_m be the network parameters of $\bar{G}_{\theta_m}$. Then, by using the chain rule, it can be obtained that

$$\begin{aligned}\nabla_{\theta_m} L_{all} &= \frac{\partial L_{all}}{\partial \bar{x}[i,j]} \cdot \frac{\partial \bar{x}[i,j]}{\partial \bar{x}_m[i,j]} \cdot \nabla_{\theta_m} \bar{G}_{\theta_m}[i,j] \\ &= \frac{\partial L_{all}}{\partial \bar{x}[i,j]} \cdot (\bar{x}_f[i,j] - \bar{x}_b[i,j]) \cdot \nabla_{\theta_m} \bar{G}_{\theta_m}[i,j].\end{aligned} \tag{14}$$

As the model converges to the second trivial solution, the foreground and background are identical, i.e. $\bar{x}_f[i,j] = \bar{x}_b[i,j]$, which causes vanishing gradients on the mask.

(Necessity) A vanishing gradient on the mask means that no gradient is passed to the mask generator. The synthesized mask is meaningless and fails to achieve foreground-background disentanglement. Therefore, the model converges to the first trivial solution when the synthesized mask is full or empty. In other cases, the model converges to the second trivial solution. □

Proof of Theorem 1: $\|\nabla_\theta L_{all}\|_2$ can be written as

$$\left\|\nabla_\theta \mathbb{E}_{z\sim p(z)}\left[\log\left(1 - D\left(G_\theta(z)\right)\right)\right]\right\|_2 \geq \sigma. \tag{15}$$

For simplicity of the following analysis, we assume the model converges to the first trivial solution and $\bar{x}_m = 1$. Under Assumption 1, it can be obtained that $\bar{x} = \bar{x}_f$ and $\partial L_{all}/\partial \bar{x}_m = \partial L_{all}/\partial \bar{x}_b = 0$. Hence, $\nabla_\theta L_{all}$ can be decomposed as $\nabla_{(\theta_\phi,\theta_f)} L_{all}$, and the gadient norms (15) can be written as:

$$\left\|\nabla_{\theta_\phi} \mathbb{E}_{z\sim p(z)}\left[\log\left(1 - D\left(\mathcal{F}(\Phi(z))\right)\right)\right]\right\|_2^2 + \left\|\nabla_{\theta_f} \mathbb{E}_{\phi\sim p(\Phi(z))}\left[\log\left(1 - D\left(\mathcal{F}_{\theta_f}(\phi)\right)\right)\right]\right\|_2^2 \geq \sigma^2. \tag{16}$$

Using Jensen's inequality and the chain rule, we show that

$$\begin{aligned}\left\|\nabla_{\theta_f} \mathbb{E}_{\phi\sim p(\Phi(z))}\left[\log\left(1 - D\left(\mathcal{F}_{\theta_f}(\phi)\right)\right)\right]\right\|_2^2 &\leq \mathbb{E}_{\phi\sim p(\Phi(z))}\left[\frac{\left\|\nabla_{\theta_f} D\left(\mathcal{F}(\phi)\right)\right\|_2^2}{\left|1 - D\left(\mathcal{F}(\phi)\right)\right|^2}\right] \\ &\leq \mathbb{E}_{\phi\sim p(\Phi(z))}\left[\frac{\left\|\nabla_{\bar{x}} D\left(\mathcal{F}(\phi)\right)\right\|_2^2 \left\|J_{\theta_f}\mathcal{F}(\phi)\right\|_2^2}{\left|1 - D\left(\mathcal{F}(\phi)\right)\right|^2}\right] \\ &< \mathbb{E}_{z\sim p(z)}\left[\frac{\left(\left\|\nabla_x D^*\left(G^*(z)\right)\right\|_2 + \epsilon\right)^2 \left\|J_{\theta_f}\mathcal{F}(\Phi(z))\right\|_2^2}{\left(\left|1 - D^*\left(G^*(z)\right)\right| - \epsilon\right)^2}\right] \\ &\leq \frac{\delta^2\epsilon^2}{(1/2-\epsilon)^2}.\end{aligned} \tag{17}$$

Then, it follows that

$$\left\|\nabla_{\theta_\phi} \mathbb{E}_{z\sim p(z)}\left[\log\left(1 - D\left(\mathcal{F}(\Phi(z))\right)\right)\right]\right\|_2^2 > \sigma^2 - \frac{\delta^2\epsilon^2}{(1/2-\epsilon)^2}. \tag{18}$$

Notice that the mask can be written as $\mathcal{M}(\Phi(z))$, which means

$$\left\|\nabla_{(\theta_\phi,\theta_m)} \mathbb{E}_{z\sim p(z)}\left[\log\left(1 - D\left(\mathcal{F}(\Phi(z))\right)\right)\right]\right\|_2^2 > \sigma^2 - \frac{\delta^2\epsilon^2}{(1/2-\epsilon)^2}. \tag{19}$$

A similar discussion can be applied to the case of $\bar{x}_m = 0$ □

Proof of Corollary 1: By Theorem 1, the larger the value of $\rho = \sigma^2 - \frac{\delta^2\epsilon^2}{(1/2-\epsilon)^2}$, the easier it is for the model to escape from trivial solutions. We observe that the δ^2 is decreased by reducing the parameter of θ_f and θ_b from the inequality $\max\{\mathbb{E}_{\phi\sim p(\Phi(z))}\|J_{\theta_f}\mathcal{F}(\phi)\|_2^2, \mathbb{E}_{\phi\sim p(\Phi(z))}\|J_{\theta_b}\mathcal{B}(\phi)\|_2^2\} \leq \delta^2$. It is notable that if the θ_f and θ_b have too few parameters or even no parameters, our model degrades into an original GAN. Therefore, the $\mathcal{F}$ and $\mathcal{B}$ should be designed as lightweight modules (e.g., adding residual connections), so as to enhance the capacity of the module with fewer parameters. Furthermore, the ρ is actually the gradient norm from the decoder Φ. It is a natural idea to increase ρ by raising the capacity of Φ. In addition, the update of Φ does not necessarily mean the update of $\mathcal{M}$. If $\mathcal{M}$ is a deep neural network, it may not be able to map the fluctuations of features to the mask space, i.e. $\mathcal{M}(\phi) \approx \mathcal{M}(\phi + \triangle\phi)$. Therefore, $\mathcal{M}$ should be designed as a shallow neural network. For example, when $\mathcal{M}$ is a sigmoid layer, it is easy to get that $\mathcal{M}(\phi) \neq \mathcal{M}(\phi + \triangle\phi)$ by the monotonicity of the sigmoid function. As for the second trivial solution, $\bar{x}_f = \bar{x}_b$ is a fragile equilibrium. We can break the identical mapping, i.e. $\mathcal{F}_{\theta_f}(\phi) = \mathcal{B}_{\theta_b}(\phi)$, by changing parameters of θ_f and θ_b or modifying the structure of the $\mathcal{F}$ and $\mathcal{B}$. □

B.2 Regularization

The binary regularization $\mathcal{L}_{\text{binary}}$ enforces the output to be close to either 0 or 1, which is as follows:

$$\mathcal{L}_{\text{binary}} = \sum_{i=1}^{N} -m_i \log_2(m_i) - (1 - m_i) * \log_2(1 - m_i), \tag{20}$$

C Additional results about ComGAN

(a) $\beta = 0$ (b) $\beta = 0.1$ (c) $\beta = 1$

Figure 9: **Visual comparisons of mask generation by fine-tuning β.**

C.1 Qualitative evaluation of mask synthesis.

We compare the visual effects of masks synthesized by ComGAN under various β. The comparison results are illustrated in Fig. 9. Without any supervised information, it is difficult for the model to distinguish which part of the image belongs to the foreground or the background. Therefore, we introduce the binary regularization βL_{binary} to separate the foreground-background regions.

(a) ComGAN (3-1,0) (b) ComGAN (3-3,0)

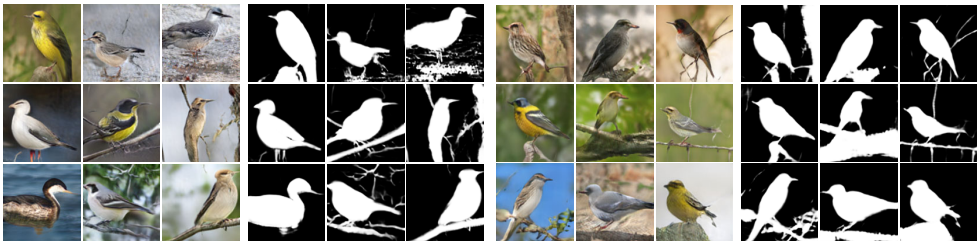
Figure 10: **Visual comparisons of ComGAN variants with imbalanced foreground and background subnetworks.** For each subfigure: (left) synthesized images, (right) corresponding masks.

C.2 Imbalance of foreground and background subnetworks

In contrast to the background, the foreground is not occluded and contains complete information. Therefore, we modify the foreground and background subnetworks with identical structures to make the two subnetworks have unbalanced network capacity. For simplicity, we define a ComGAN variant called ComGAN ($N_F - N_B$, N_M), where N_F, N_B and N_M denote the number of ResBlocks in the foreground, background and mask subnetworks, respectively. As shown in Fig. 10, a sizeable visual gap exists between the masks produced by ComGAN(3-1,0) and ComGAN (3-3,0). The masks in Fig. 10(a) contain more foreground information, such as branches and rocks, even if the information is insignificant. In contrast, the masks Fig. 10(b) lack partial details of objects, especially the bird's feet. It seems that we obtain higher-quality segmentation masks by trade-offs between the capacity of the foreground and background subnetworks.

C.3 Relaxing constraints on mask subnetwork

Corollary 1 requires that the $\mathcal{M}$ is a shallow network. The property is a strong constraint on the network design. Therefore, we show that a ComGAN variant avoids trivial solutions even when



(a) ComGAN (3-3,1) (b) ComGAN (3-3,2)

Figure 11: **Visual comparisons of ComGAN variants with various mask subnetworks.** For each subfigure: (left) synthesized images, (right) corresponding masks.

Methods	Sup.	FID ↓	IS↑
Triple-GAN [49]	Semi.	140.94	3.94±0.06
EnhancedTGAN [40]	Semi.	133.57	4.17±0.06
Triangle-GAN [50]	Semi.	96.42	4.36±0.05
R^3-CGAN[51]	Semi.	88.62	4.43±0.06
SSC-GAN§ [27]	Semi.	**20.03**	4.68±0.04
FineGAN§[25]	Weak.	46.68	4.62±0.03
MixNMatch§ [31]	Weak.	45.59	**4.78±0.08**
SN-GAN [52]	Unsup.	160.09	4.21±0.05
ComGAN(3-3,0)	Unsup.	21.72	4.73±0.08
ComGAN(3-3,1)	Unsup.	20.41	4.76±0.11
ComGAN(3-3,2)	Unsup.	20.80	4.68±0.07

Table 5: **Image synthesis results on CUB.** ComGAN variants are compared with the state-of-the-art GAN-based models in image synthesis.

Methods	IoU↑	Dice↑
W-Net [34]	24.8	38.9
GrabCut [53]	30.2	42.7
ReDO†* [35]	46.5	60.2
OneGAN◇* [21]	55.5	69.2
IODINE† [54]	30.9	44.6
Slot-Attn.† [55]	35.6	51.5
IEM+SegNet [36]	55.1	68.7
DRC [44]	56.4	**70.9**
ComGAN (3-3,0)	50.1	64.6
ComGAN (3-3,1)	**56.9**	69.8
ComGAN (3-3,2)	53.9	67.5

Table 6: **Segmentation results on CUB.** ComGAN variants are compared with the state-of-the-art un(weakly-)supervised segmentation methods.

the mask subnetwork is not a sigmoid function. ComGAN variants are designed by increasing the number of ResBlocks in the mask subnetwork. It can be seen from Fig. 10(b) and Fig 11 that the extension of the mask subnetwork does not result in trivial solutions and provides masks that retain more details.

C.4 Quantitative evaluation of image generation.

As summarized in Table 5, ComGAN variants can achieve comparable FID scores. We conjecture that improved image synthesis quality is attributable to ComGAN's network architecture. The mask subnetwork and binary regularization force foreground and background subnetworks to focus on learning simpler and continuous foreground and background regions. Furthermore, the experimental results on GAN-based image composition models such as FineGAN and MixNMatch indicate that trivial solutions degrade the quality of image generation. Specifically, the first trivial solution degrades the generation process of composing the foreground and background to a single generation. The second trivial solution induces interference between the foreground and background, decreasing the composite image quality.

C.5 Quantitative evaluation of segmentation masks.

To highlight that ComGAN variants generate high semantic masks, we obtain the predicted segmentation masks by the method in Section 3.3. As summarized in Table 6, ComGAN variants are simple and exhibit comparable performance with only two regularizations. Possibly limited by the known drawbacks of GANs such as unstable training, ComGAN variants have differences in segmentation results. Moreover, we observe that the quality of the predicted segmentation masks is improved when the image distribution is closer (i.e., the lower FID scores).

	N=100	N=150	N= 200	N=250	N=300
FID ↓	16.94	17.33	**16.26**	16.71	17.39
IS ↑	4.63	4.71	**4.79**	4.76	4.69

Table 7: **Robustness results for variable dimensions.**

D Additional discussion about DS-ComGAN

D.1 Influence of dimension of variables

Table 7 shows the FID and IS of DS-ComGAN trained on the CUD by varying the dimension of variables N. We observe that our model achieves comparable performance with various N. As discussed in section 3.2, the N can be considered as the number of categories in clusters. Therefore, when N equals 200, model performance is marginally improved.

D.2 Ablation study on global feature extraction

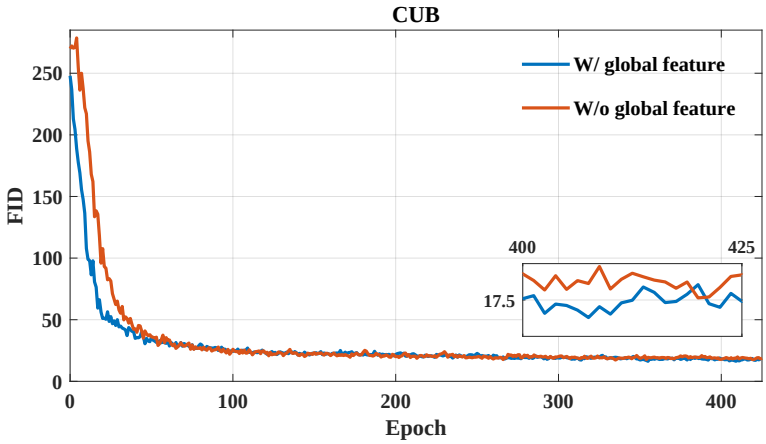


Figure 12: **The FID curves of training DS-ComGAN w/ or w/o global feature extraction.**

As shown in Fig. 12, the model performs global feature extraction on the CUB dataset, which leads to a rapid convergence and decrease of 1 to 2 points in the FID scores. Of note is that the extraction of global features consumes a large amount of GPU video memory and slightly improves the quality of image generation. However, the operation is necessary for specific datasets such as Stanford Cars, because it permits the mask to capture the entire distinction between the foreground and background, as illustrated in Fig. 13.

(a) Image synthesis w/ global feature (b) Image synthesis w/o global feature

Figure 13: **Ablation study on extraction of global features.** For each subfigure: (left) synthesized images, (right) corresponding masks.

D.3 Additional unsupervised segmentation results

See Fig. 14, Fig. 15, Fig. 16, Fig. 17, and Fig. 18 for more unsupervised segmentation results by our model on each dataset. From top to bottom, we display: (i) observed images, (ii) ground-truth masks, and (iii) predicted masks in each figure.

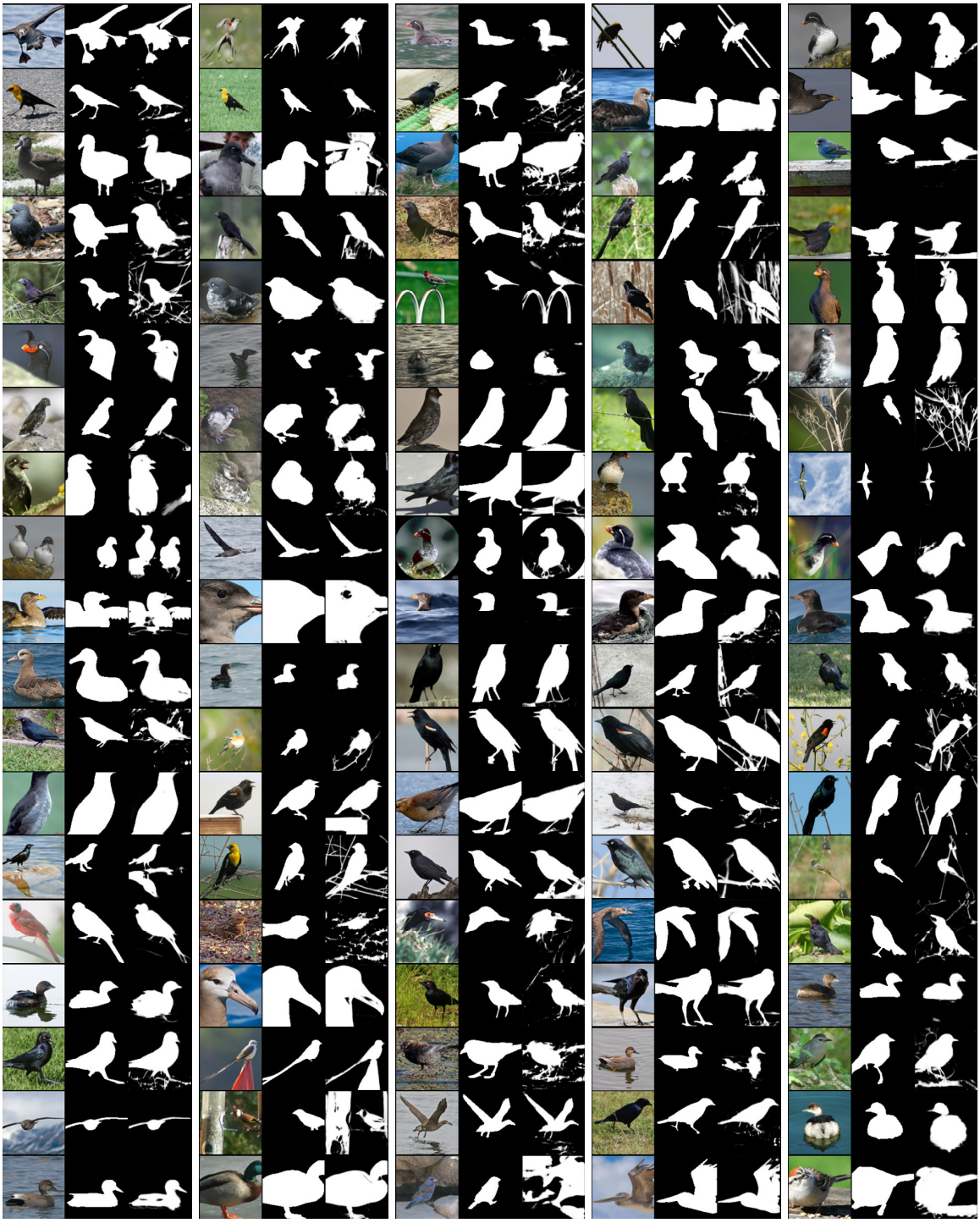


Figure 14: More randomly sampled segmentation masks on CUB.

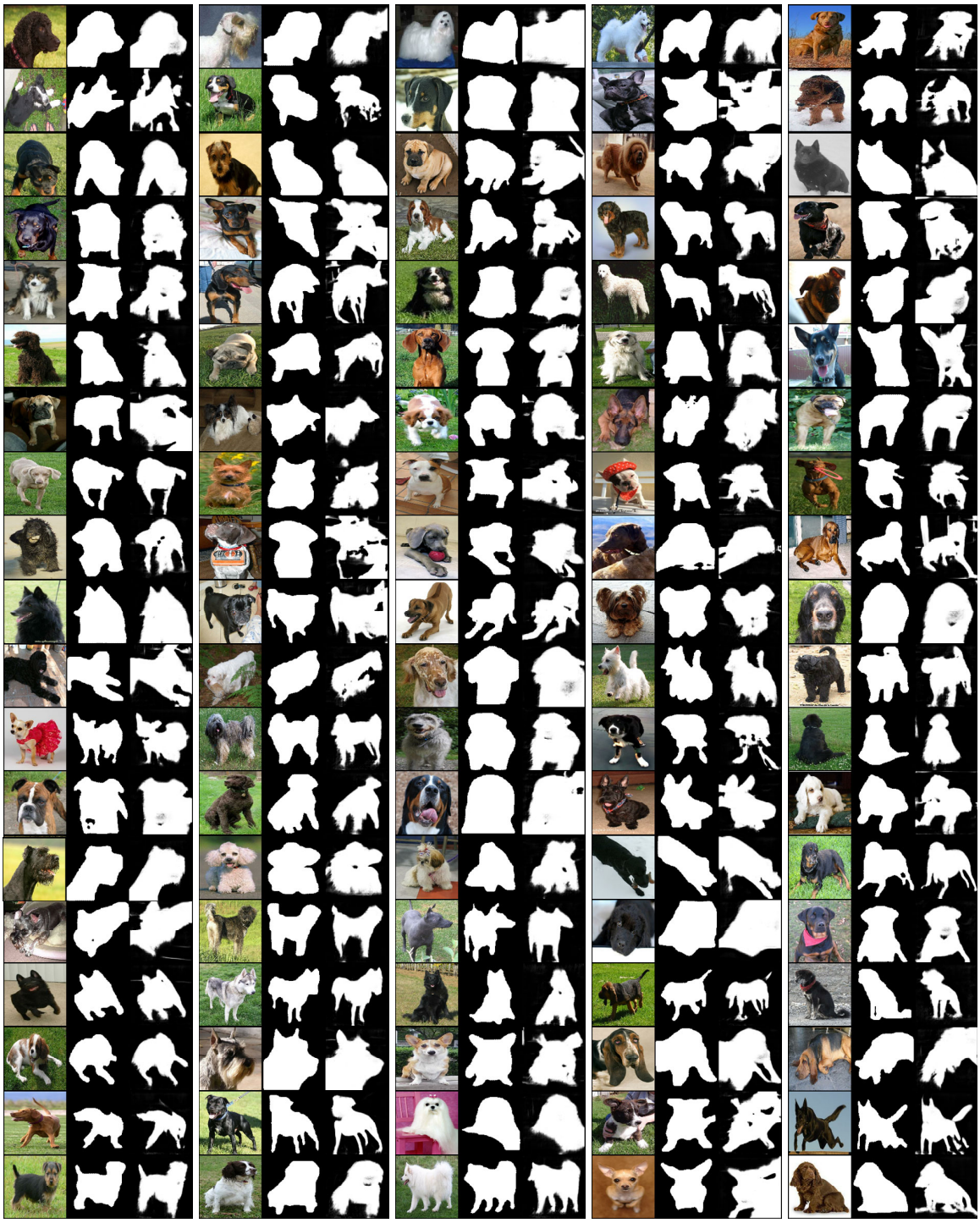


Figure 15: More randomly sampled segmentation masks on Stanford Dogs.

Figure 16: More randomly sampled segmentation masks on Stanford Cars.

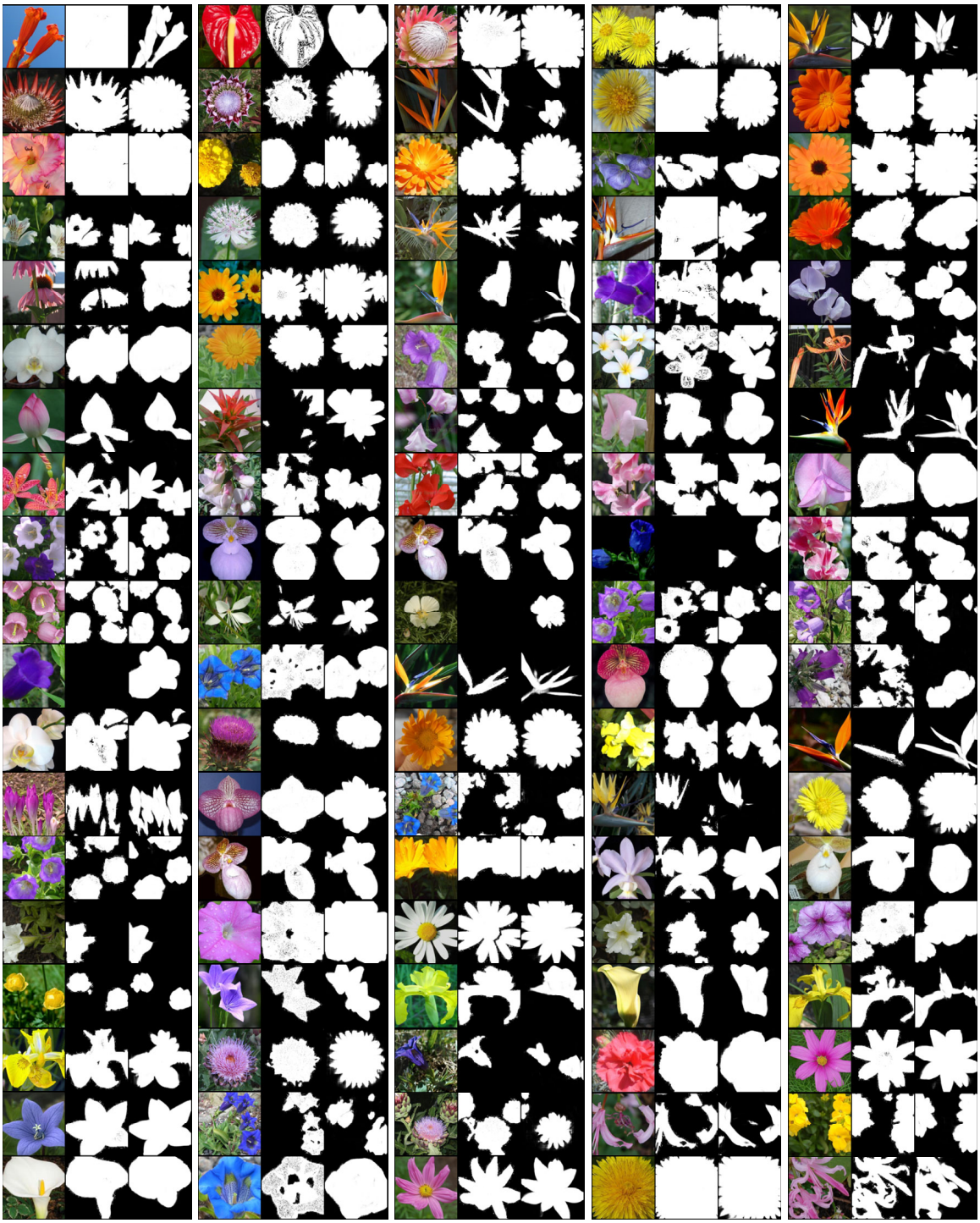


Figure 17: More randomly sampled segmentation masks on Flowers.

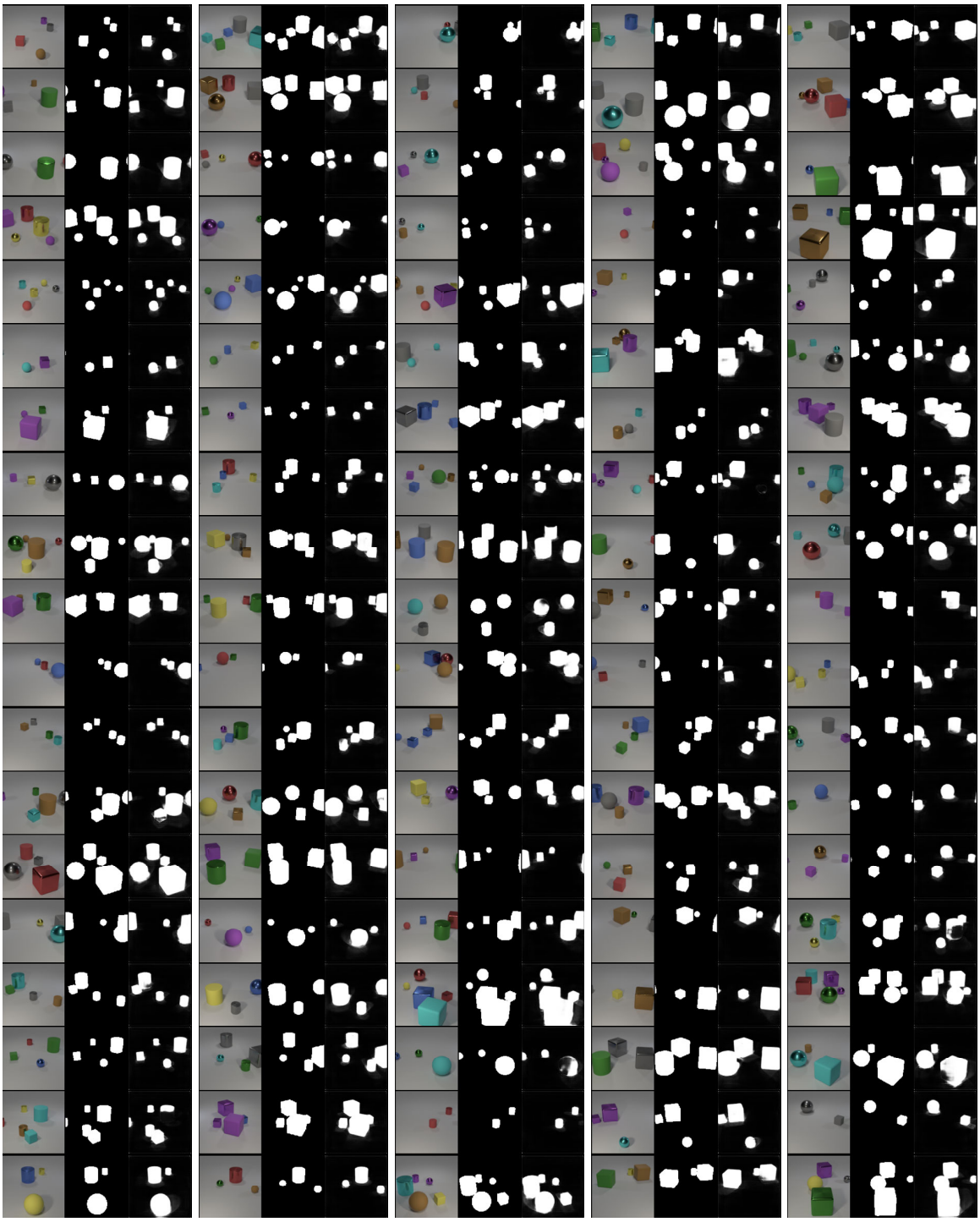


Figure 18: More randomly sampled segmentation masks on CLEVR6.