# ComGAN: Unsupervised Disentanglement and Segmentation via Image Composition

**Anonymous Author(s)**
Affiliation
Address
email

## 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. Our basic idea aims to generate the foreground, background and mask similar to the true data distribution. 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 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 baselines 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.[1]

## 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 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

[1] Code and data are available at https://anonymous.4open.science/r/ComGAN-D174

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

work alleviates 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 the foreground, background and mask similar to the true data distribution. In addition, ComGAN effectively avoids trivial solutions by the network architecture, which does not require any supervised information and explicit regularization.

To highlight the flexibility of ComGAN, we generalize it to two related tasks, namely image disentanglement, and unsupervised segmentation. For the image disentanglement task, a series of works [25, 26, 27, 28] hierarchically generates images and disentangles the background, the object's shape and its appearance by bounding box annotation. 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. SSC-GAN [29] 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. These methods rely on additional supervised information to achieve multi-factor disentanglement. Recent work on unsupervised segmentation. ReDO (ReDrawing of Objects) [30] captures the mask of the object, relying on the assumption that changing the foreground objects does not affect the overall data distribution. Further, to better perform unsupervised segmentation and detection in video, Yang et al. [14, 15] propose a stringent assumption that the optical flows in foregrounds and backgrounds are independent. A similar assumption is implemented in the method [31], which performs unsupervised segmentation by partitioning images into maximally independent sets. Labels4Free [12] performs unsupervised object segmentation based on the idea that exact coupling between foreground and background is highly non-trivial. Despite these methods achieving decent performance on unsupervised segmentation, stringent assumptions limit their applicability. In this paper, to confront two challenges mentioned above, we redesign two fully unsupervised modules based on ComGAN, i.e., the disentanglement module and the segmentation module, and then rename the whole network DS-ComGAN. DS-ComGAN relies only on mild assumption 1 to achieve image disentanglement and object segmentation in a fully unsupervised manner. 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 network architecture perspective and achieve foreground-background disentanglement with only an adversarial loss and a binary regularization.

2. We propose DS-ComGAN, a redesigned network based on ComGAN, which achieves image disentanglement and object segmentation in a fully unsupervised manner. Experiments show that DS-ComGAN exhibits excellent performance on both tasks, even outperforming semi-supervised and weakly supervised baselines.

## 2 Related Work

**GAN-based image composition.** Image composition can be regarded as combining multiple visual areas to construct a realistic image [32]. 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 [33] proposes semantic image generation control by imposing regularization of mutual information. These methods [34, 35] learn the independent latent characteristics of an object, especially its appearance and pose. MixNMatch [27] extends the model [25], which disentangles and encodes four variable factors (background, object pose, shape, and texture) from real images. Along this direction, PartGAN [26] achieves part-level decomposition by learning a part generator, and OneGAN [21] simultaneously solves multiple tasks such as image disentanglement and unsupervised segmentation. To deal with complex real-world data, CGN [19] decomposes the image generation process into independent causal mechanisms, and allows for generating counterfactual images. SSG-GAN [29] uses a small amount of labeled data and controls the image generation process by class-independent, cross-class, and class variables. Although these methods exhibit strong performance, they all rely on additional supervised information to assist the model in identifying foregrounds and backgrounds.

**Unsupervised Object Segmentation.** Unsupervised object segmentation aims to extract a useful interpretation from an image in an unsupervised way. W-net [36] extracts foreground objects through an encoder-decoder framework and minimizes reconstruction errors. ReDO [30] learns segment images by replacing any segmented object in a scene without affecting the image realism. 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. Specifically, IEM [31] 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.

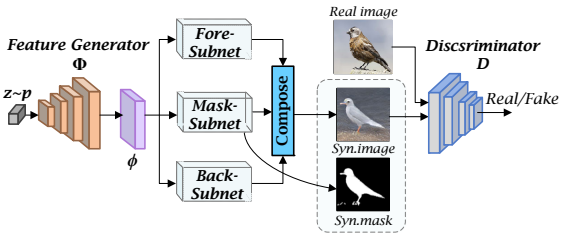
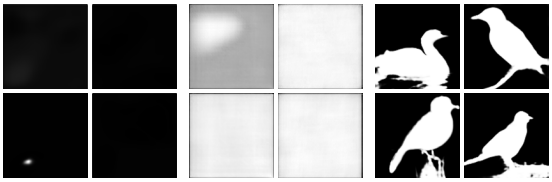



Figure 1: **Overview of ComGAN.** The generator of ComGAN is composed of the feature generator $\Phi$ and three independent subnetworks. Three independent subnetworks synthesize the foreground, background and mask with features $\phi$ as a shared input. The 'compose' indicates the utilization of assumption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



(a) Model $\Pi_1$ (b) Model $\Pi_2$ (c) ComGAN

Figure 2: **Comparison results of the synthesized mask.** ComGAN synthesizes high semantic masks, while the masks synthesized by models $\Pi_1$ and $\Pi_2$ both converge to trivial solutions. Note that $\Pi_1$ has the same structure as [25, 26, 27, 28], and $\Pi_2$ has a similar structure to [11, 19]. All models include identical adversarial loss and binary regularization $\beta L_{binary}$ from [19]. For each model: (left) $\beta$ = 0.1, (right) $\beta$= 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 realistic an image $\bar{x}$ utilizing assumption 1. If there exist vanishing gradients on the mask, i.e., $\partial L_{all}/\partial \bar{x}_m = 0$, then the model learns the trivial solutions.*

*Proof:* Under assumption 1, the $\bar{x}$ at each pixel $i, j \in \mathbb{R}$ can be written as follows:

$$\bar{x}[i,j] = \bar{x}_f[i,j] \cdot \bar{x}_m[i,j] + \bar{x}_b[i,j] \cdot (1 - \bar{x}_m[i,j]), \tag{2}$$

where $\bar{x}_f$ is the foreground and $\bar{x}_b$ is the background. Let $\bar{G}_{\theta_m}$ be masks generator and $\theta_m$ be the network parameters of $\bar{G}_{\theta_m}$. Then, by using the chain rule, it can be obtained that

$$\begin{aligned}\nabla_{\theta_m} L_{all} &= \frac{\partial L_{all}}{\partial \bar{x}[i,j]} \cdot \frac{\partial \bar{x}[i,j]}{\partial \bar{x}_m[i,j]} \cdot \nabla_{\theta_m} \bar{G}_{\theta_m}[i,j] \\ &= \frac{\partial L_{all}}{\partial \bar{x}[i,j]} \cdot (\bar{x}_f[i,j] - \bar{x}_b[i,j]) \cdot \nabla_{\theta_m} \bar{G}_{\theta_m}[i,j].\end{aligned} \tag{3}$$

As observed in equation (3), two kinds of special points both cause vanishing gradients on the mask, i.e., $\partial L_{all}/\partial \bar{x}_m = 0$. The first special point is $\bar{x}^*_m[i,j] = 1$ or 0. As $\bar{x}^*_m[i,j]$ accumulate, the masks converge to the first trivial solution. The second special point is based on an equilibrium, i.e., $\bar{x}^*_f[i,j] = \bar{x}^*_b[i,j]$. Similarly, the model learns a second trivial solution. □

**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 $\Phi$ 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 defined as $\mathcal{F}$, $\mathcal{B}$ and $\mathcal{M}$ synthesize the foreground, background and mask, respectively. In this paper, $\mathcal{F}$ and $\mathcal{B}$ both are composed of two convolution layer and two consecutive residual blocks and $\mathcal{M}$ includes a sigmod layer. Afterwards, the foreground, background and mask are defined as: $\bar{x}_f = \mathcal{F}(\phi)$, $\bar{x}_b = \mathcal{B}(\phi)$ and $\bar{x}_m = \mathcal{M}(\phi)$. Driven by assumption 1, the composite image $\bar{x}$ is defined as:

$$\bar{x} = \mathcal{F}(G(z)) \odot \mathcal{M}(G(z)) + \mathcal{B}(G(z)) \odot (1 - \mathcal{M}(G(z))). \tag{4}$$

Then, the adversarial training loss $L_D^{adv}$ is defined as follows:

$$\min_{\Phi,\mathcal{F},\mathcal{B},\mathcal{M}} \max_{D} \mathcal{L}_D^{adv} = \mathbb{E}_{x\sim p_{\text{data}}}\left[\log\left(1 - D(x)\right)\right)\right] + \mathbb{E}_{z\sim p_0}\left[\log D(\bar{x})\right]. \tag{5}$$

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

**Proposition 1** *Under assumption 1, if a model that synthesizes realistic images has a network architecture similar to ComGAN, then the model avoids trivial solutions and generates the foreground, background, and mask similar to the true data distribution.*

*Proof:* Under assumption 1, we decompose the real image $x$ into foreground $x_f$, background $x_b$ and mask $x_m$, and obtain their distribution as follows:

$$\begin{aligned}\mathbb{E}_{x\sim p_{data}}(x) =& \mathbb{E}_{x_m\sim p(x_m)} \odot [\mathbb{E}_{x'_f\sim p(x_f|x_m)}(x'_f \odot x_m)] \\ &+ \mathbb{E}_{x_m\sim p(x_m)} \odot [\mathbb{E}_{x'_b\sim p(x_b|x_m)}(x'_b \odot (1 - x_m))],\end{aligned} \tag{6}$$

where $p_{data}$ is the true data distribution. Since $\mathcal{M}$ is a sigmod function, it follows that $\bar{x}_m = \mathcal{M}^{-1}(G(z))$ from equation (4). Similarly, we obtain the foreground, background and mask distribution synthesized by the model.

$$\begin{aligned}\mathbb{E}_{z\sim p_0}(\bar{x}) =& \mathbb{E}_{\bar{x}_m\sim \mathcal{M}(G(z))} \odot [\mathbb{E}_{\bar{x}'_f\sim \mathcal{F}(\mathcal{M}^{-1}(\bar{x}_m))}(\bar{x}'_f \odot \bar{x}_m)] \\ &+ \mathbb{E}_{\bar{x}_m\sim \mathcal{M}(G(z))} \odot [\mathbb{E}_{\bar{x}'_b\sim \mathcal{B}(\mathcal{M}^{-1}(\bar{x}_m))}(\bar{x}'_b \odot (1 - \bar{x}_m))]\end{aligned} \tag{7}$$

Two conclusions are drawn from equations (6) and (7). First, the model generates the foreground, background and mask similar to the real data distribution, i.e., $\mathcal{M}(G(z))$, $\mathcal{F}(\mathcal{M}^{-1}(\bar{x}_m))$ and $\mathcal{B}(\mathcal{M}^{-1}(\bar{x}_m))$ approximate distribution of $p(x_m)$, $p(x_f|x_m)$ and $p(x_b|x_m)$. In other words, the model imitates and learns the internal correlation between $x_f$, $x_b$ and $x_m$ from the network architecture perspective. Second, the model can effectively solve trivial solutions. From equation (7), it is clear that any change of foreground or background affects the mask, which means that the gradients from the $x_f$ or $x_b$ are passed to the $x_m$, i.e., $\nabla_{\theta_m} \bar{G}_{\theta_m} \neq 0$ when $\partial L_{all}/\partial \bar{x}_m = 0$. □

### 3.2 The Disentanglement Module of DS-ComGAN

Image generation is controlled by a mask variable $z$, a foreground variable $f$ and a background variable $b$, where $f$ and $b$ are both $N$ dimensional random one-hot vectors. Similarly to ComGAN,

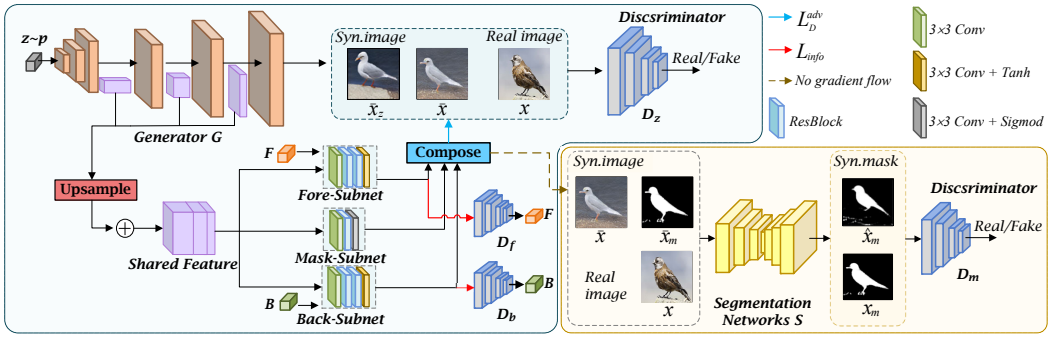



Figure 3: **Overview of DS-ComGAN.** DS-ComGAN consists of two modules: (i) the disentanglement module (blue); (ii) the segmentation model (yellow). The generator $G$ synthesizes $\bar{x}_z$ to perform adversarial training with the discriminator $D_z$ for $G$ to contain more global features. Shared features 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 can synthesize high semantic masks, the segmentation module performs unsupervised segmentation with image distribution alignment and mask distribution alignment via adversarial training.

the generator of DS-ComGAN 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 features for shared network.** 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. In the first step, we synthesize realistic images, i.e., $\bar{x}_z = G(z)$, and then extract the activation map from the last three layers inside $G$, denoted as $a_1$, $a_2$ and $a_3$. Then, shared features with global information can be written as follows:

$$\phi = \mathbb{U}(a_1) \oplus \mathbb{U}(a_2) \oplus \mathbb{U}(a_3) \tag{8}$$

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:

$$\bar{x} = \mathcal{F}(\phi, f) \odot \mathcal{M}(\phi) + \mathcal{B}(\phi, b) \odot (1 - \mathcal{M}(\phi)). \tag{9}$$

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:

$$\min_{G,\mathcal{F},\mathcal{B},\mathcal{M}} \max_{D_z} \mathcal{L}_{D_z}^{adv} = \mathbb{E}_x \left[\log D_z(x)\right] + \mathbb{E}_z \left[\log(1 - D_z(\bar{x}_z))\right] + \mathbb{E}_{z,f,b} \left[\log(1 - D_z(\bar{x}))\right]. \tag{10}$$

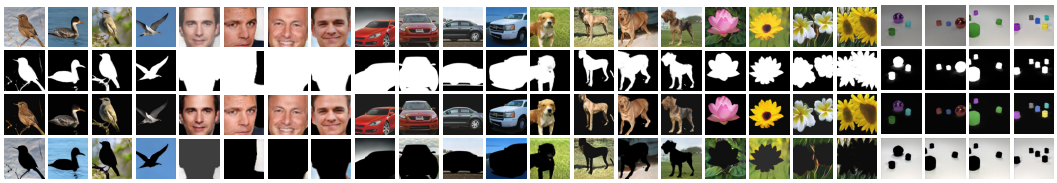
**Controllable regularization for foreground-background subnetworks.** A mutual information regularization is incorporated into the $\mathcal{F}$ and $\mathcal{B}$ to associate the variables $f$ and $b$ with the synthesized $\bar{x}_f$ and $\bar{x}_b$. Similar to InfoGAN [33], we obtain similar results by maximizing the lower bound on the variance of the mutual information.

$$\max_{D_f, D_b} L_{info} = I(\bar{x}_f, f) + I(\bar{x}_b, b) \geq \mathbb{E}_{z,f,b}[\log D_f(f \mid \bar{x}_f) + \log D_b(b \mid \bar{x}_b)], \tag{11}$$

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 foreground 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. Unlike previous GAN-based unsupervised segmentation methods [31, 11, 12, 30, 18], our method is not limited by



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

Figure 4: **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.

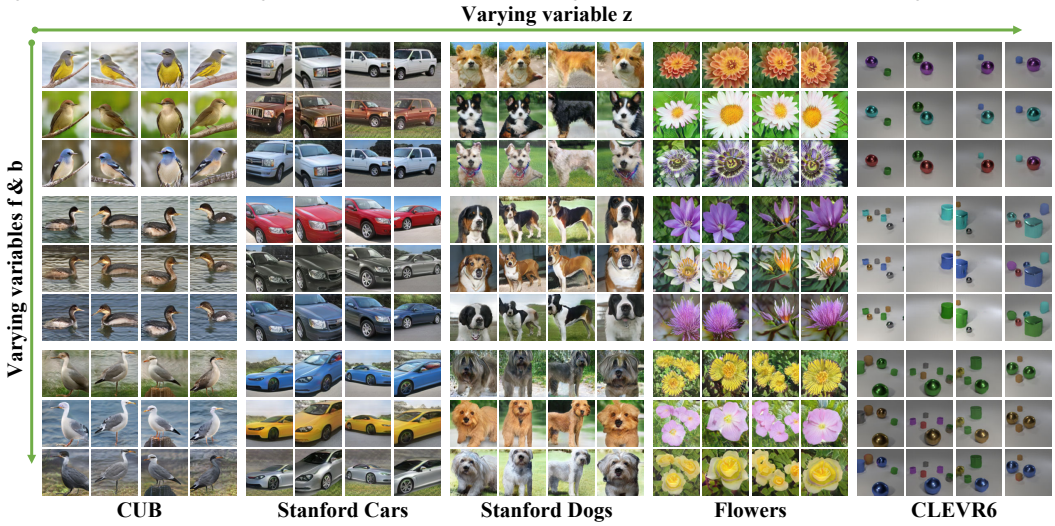



Figure 5: **The controllable image generation via varying variables $f$, $b$ and $z$.** Zoom in for better visibility.

the stringent assumption that the foreground and background are largely independent. Our key idea is to exploit the high semantic masks synthesized by our model 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.** The first step is to obtain the masks of $x$ and $\bar{x}$, i.e., $x_m = \mathcal{S}(x)$ and $\hat{x}_m = \mathcal{S}(\bar{x})$. Similarly, we align mask distribution among $x_m$, $\hat{x}_m$ and $\bar{x}_m$ via adversarial training,

$$\min_{\mathcal{S}} \max_{D_m} \mathcal{L}_{D_m}^{adv} = \mathbb{E}_{\bar{x}_m, \hat{x}_m, x_m}[\log D_m(\bar{x}_m) + \log(1 - D_m(\hat{x}_m)) + \log(1 - D_m(x_m))]. \quad (12)$$

To ensure that the masks predicted by the segmentation network are consistent with the input image, rather than merely learning the mask distribution. Hence, the corresponding consistency loss $L_{cons}$ is as follows:

$$\min_{\mathcal{S}} \mathcal{L}_{cons} = \lambda \cdot \mathbb{E}_{\bar{x}, \bar{x}_m} \left[ \| \bar{x}_m - \mathcal{S}(\bar{x}) \| \right], \quad (13)$$

where $\lambda$ is the weight of $L_{cons}$. We transform the unsupervised segmentation task into two adversarial tasks and acquire segmentation masks by enhancing image and mask distribution consistency.

# 4 Experiments

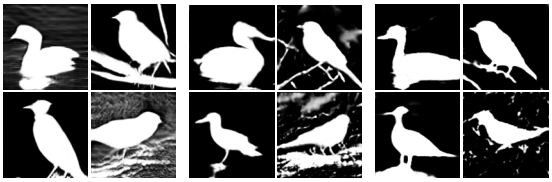
## 4.1 Experimental Setup

**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.

**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 Mask synthesis based on ComGAN

**Avoidance of trivial solutions.** To demonstrate that ComGAN efficiently solves trivial solutions, we selected two typical GAN-based image compositional models $\Pi_1$ and $\Pi_2$ for comparison. All models include identical adversarial loss and binary regularization $\beta L_{binary}$ from [19], and synthesize realistic images by the assumption 1. The model $\Pi_1$ is obtained from FineGAN [25]. $\Pi_1$ consists of two generators $G_1$ and $G_2$, where $G_1$ generates foreground and mask and $G_2$ generates background. The model $\Pi_2$ is similar to PerturbGAN [11] and CGN [19]. $\Pi_2$ 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; (2) we simplify BigGAN [48] and StyleGAN [49] with SimpleGAN in $\Pi_2$ and retrain it. As shown in Fig. 2, the mask synthesized by ComGAN avoids convergence to trivial solutions and has non-trivial semantic information.



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

Figure 6: **Visual comparisons of mask generation under various binarization.**

(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 1: **The results of the baseline and variants on CUB.**

| | $N$=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 2: **Robustness results of ComGAN with various dimension.**

**Qualitative evaluation of mask synthesis.** We compare the visual effects of masks synthesized by ComGAN under various binarization. The comparison results are illustrated in Fig. 6. 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 $\beta L_{binary}$ [19] to separate the foreground-background regions.

### 4.3 Image Disentanglement based on DS-ComGAN

**Controllable image disentanglement of independent variables.** Fig. 5 illustrates that by varying three independent variables, DS-ComGAN synthesizes realistic and diverse images such that the model achieves controlled image disentanglement. We see that variable $z$ is associated with the masks. 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.

**Quantitative evaluation of image generation.** We compare DS-ComGAN with state-of-the-art GAN-based generative models without any advanced GAN training strategy. As illustrated in Table 3, the results show that DS-ComGAN significantly outperforms all unsupervised and weakly supervised baselines, and even outperforms semi-supervised baseline [29] that requires additional label as input. FineGAN and MixNMatch may be limited by the conflict of foreground and background generators,

| Methods | Sup. | CUB | | FS-100 | | Stanford-Cars | |
|---|---|---|---|---|---|---|---|
| | | FID ↓ | IS↑ | FID ↓ | IS↑ | FID ↓ | IS↑ |
| Triple-GAN [50] | 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 [51] | Semi. | 96.42 | 4.36±0.05 | 35.49 | 1.71±0.04 | 61.44 | 2.77±0.10 |
| $R^3$-CGAN [52] | Semi. | 88.62 | 4.43±0.06 | 25.28 | 1.73±0.02 | 44.57 | 3.05±0.04 |
| SSC-GAN§ [29] | 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§ [27] | Weak. | 45.59 | 4.78±0.08 | 25.63 | 1.71±0.05 | 45.94 | 2.60±0.05 |
| SN-GAN [53] | 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 3: **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.

such as trivial solutions. 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 DS-ComGAN on diverse datasets. Stanford-Dogs and Flowers datasets typically include a single foreground object, whereas CLEVR6 contains a variety of occluded and truncated objects. Table 4 shows that DS-ComGAN outperforms state-of-the-art GAN-based models in the task of image disentanglement.

| Methods | Single Object | | | | Multi-Object | |
|---|---|---|---|---|---|---|
| | Stanfor-Dogs | | Flowers | | CLEVR6 | |
| | FID ↓ | IS↑ | FID ↓ | IS↑ | FID ↓ | IS↑ |
| SSC-GAN§ [29] | 64.26 | 8.97±0.12 | 29.09 | 3.41±0.03 | \ | \ |
| FineGAN§ [25] | 69.52 | 8.27±0.17 | \ | \ | \ | \ |
| MixNMatch§ [27] | 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 4: **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.

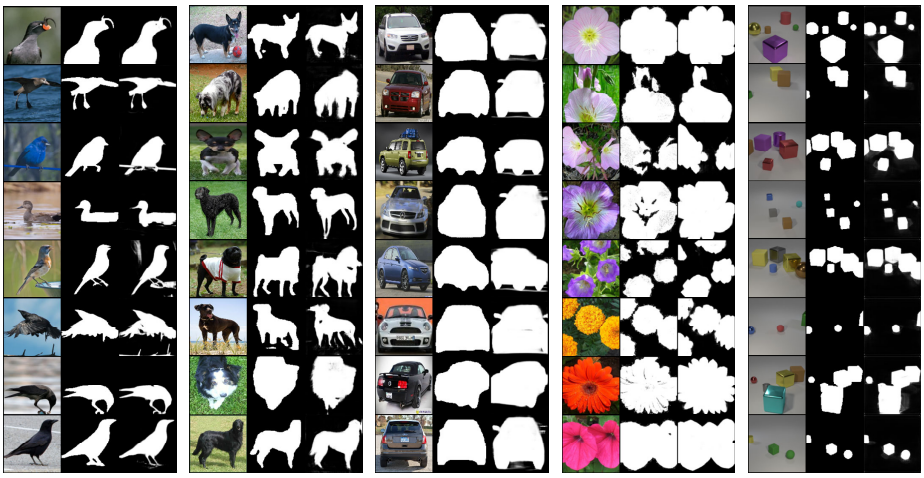
**Effectiveness of model components** To evaluate the contribution of each proposed component, we analyze quantitatively the performance margin between the baseline model and ComGAN. As demonstrated in Table 1, the incorporation of $\mathcal{F}$, $\mathcal{B}$ and $\mathcal{M}$ significantly improves 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.

**Influence of dimension of variables.** Table 2 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.

### 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. 5, 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 fine-tuning $\lambda$.



(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.

| | Single Object | | | | | | | | Multi-Object | |
|---|---|---|---|---|---|---|---|---|---|---|
| | CUB | | Stanford-Dogs | | Stanford-Cars | | Flowers | | CLEVR6 | |
| Methods | IoU↑ | Dice↑ | IoU↑ | Dice↑ | IoU↑ | Dice↑ | IoU↑ | Dice↑ | IoU↑ | Dice↑ |
| W-Net [36] | 24.8 | 38.9 | 47.7 | 62.1 | 52.8 | 67.6 | - | - | - | - |
| GrabCut [54] | 30.2 | 42.7 | 58.3 | 70.9 | 61.3 | 73.1 | 69.2 | 79.1 | 19.0 | 30.5 |
| ReDO†⋇ [30] | 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† [55] | 30.9 | 44.6 | 54.4 | 67.0 | 51.7 | 67.3 | - | - | 19.9 | 32.4 |
| Slot-Attn.† [56] | 35.6 | 51.5 | 38.6 | 55.3 | 41.3 | 58.3 | - | - | 83.6 | 90.7 |
| IEM+SegNet [31] | 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 5: **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.

**Quantitative comparisons to prior work.** Table 5 demonstrates that DS-ComGAN outperforms all unsupervised and weakly unsupervised baselines significantly. The result benefits from the high semantic masks synthesized by DS-ComGAN. In other words, based on the ComGAN's network architecture, DS-ComGAN generates masks similar to the true data distribution, which reduces the noise from regularization and trivial solutions.

## 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 [57] ). We conjecture that the limitation might relate to the capacity of model. More discussions can be found in the supplementary material.

We first analyze that the source of the two kinds of trivial solutions is vanishing gradients on the mask. Subsequently, we propose ComGAN, a simple unsupervised generative model, which generates the foreground, background and mask similar to true data distribution 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.

# 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] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

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

[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] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. In *International Conference on Learning Representations (ICLR)*, 2018.

[49] 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.

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

[51] 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.

[52] 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.

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

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

[55] 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.

[56] 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.

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

# 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 Lemma 1, Proposition 1.
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://anonymous.4open.science/r/ComGAN-D174`. 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]