Learning Foreground-Background Segmentation from Improved Layered GANs

Yu Yang¹, Hakan Bilen², Qiran Zou¹, Wing Yin Cheung¹, Xiangyang Ji¹
¹ Tsinghua University, BNRist ² The University of Edinburgh
yang-yu16@mails.tsinghua.edu.cn, hbilen@ed.ac.uk,
{zouqr19,zhangyx20}@mails.tsinghua.edu.cn, xyji@tsinghua.edu.cn

Abstract

Deep learning approaches heavily rely on high-quality human supervision which is nonetheless expensive, time-consuming, and error-prone, especially for image segmentation task. In this paper, we propose a method to automatically synthesize paired photo-realistic images and segmentation masks for the use of training a foreground-background segmentation network. In particular, we learn a generative adversarial network that decomposes an image into foreground and background layers, and avoid trivial decompositions by maximizing mutual information between generated images and latent variables. The improved layered GANs can synthesize higher quality datasets from which segmentation networks of higher performance can be learned. Moreover, the segmentation networks are employed to stabilize the training of layered GANs in return, which are further alternately trained with Layered GANs. Experiments on a variety of single-object datasets show that our method achieves competitive generation quality and segmentation performance compared to related methods.

1. Introduction

Deep learning approaches continue to dramatically push the state-of-the-art in most supervised computer vision tasks, but they demand tremendous number of human annotations which are expensive, time-consuming, and error-prone, especially for image segmentation task. Hence there is a growing interest in reducing the reliance on human supervision. Towards this target, we investigate learning to synthesize paired images and segmentation masks without segmentation annotation for the use of training segmentation network (Fig. 1).

In the recent years, generative adversarial networks (GANs) [18] repeatedly show its ability to not only synthesize highly photorealistic images [26, 27, 7] but also disentangle attributes [26, 27] and geometric factors [54]. To make GANs factorize image layers (i.e. foreground and background), a body of works [52, 47, 5, 10, 6], dubbed layered GANs in this paper, employ a compositional generation process which first generates background and foreground objects and then composes them to obtain the synthetic image. This compositional generation process, however, does not guarantee an expected factorization and is prone to trivial compositions such as “all as foreground” and “all as background” (Fig. 2(a) and Fig. 4) without any supervision or any regularization.

PerturbGAN [6] addresses this issue by perturbing the foreground position during composition, as such perturbation leads to inferior generated images for all-as-foreground decomposition and maintains decent generated images for expected decomposition (Fig. 2(b)). Despite so, perturbation is unable to solve the all-as-background issue, or empty foreground issue, where it can not affect the fidelity of generated image. To alleviate this problem, PerturbGAN [6] penalizes the cases where the number of foreground pixels are smaller than a threshold. However, this method is not always effective and the additionally introduced hyperparameter is sensitive to object scale, object category and dataset.

We propose a GAN that synthesizes background and foreground from a pair of latent codes. In more details, one of the latent codes, dubbed public code, is shared across
background and foreground generation, while the other, dubbed private code, is only reserved for foreground generation. In contrast to PerturbGAN [6], we not only employ perturbation and require our generators to generate photorealistic images but also diversify the generated foreground by maximizing the mutual information between the composed image and the private code as well as the mutual information between the foreground mask and the private code. To this end, the all-as-background issue can be effectively mitigated and the layered GANs decompose foreground and background in a superior way.

Our method is reminiscent of a rich body of work [16, 8, 19, 15, 34, 14, 3] on object-centric scene generation, which is mostly based on variational autoencoders (VAEs). These models are able to decompose and generate multi-instance complex scenes. Nonetheless, these methods are mostly demonstrated on synthetic scene but seldom applied to real-world images [14]. In contrast, our method has an advantage in dealing with realistic images but lack the ability to model multi-instance generation.

Our method presents a solution to unsupervised foreground-background segmentation (Fig. 1) where layered GANs are trained to generated synthetic training data from which segmentation networks are trained. In return, the segmentation networks are used to further regularize the generation of layered GANs. These two steps are alternated which finally leads to a high-performance segmentation network. Our method learns segmentation with the analysis-by-synthesis principle, which is different from earlier work [44, 21, 25, 2, 9] that learns to discriminate foreground and background pixels using shallow classifiers on hand-crafted features and the recent trend to simultaneously learn per-pixel representations and clustering [24, 40, 29] with deep neural networks.

We evaluate our method on a variety of single-object datasets including Caltech-UCSD Birds 200-2011 (CUB), Stanford Cars, Stanford Dogs, and Amazon Picking Challenge (APC), where it outperforms the baseline methods and achieves competitive performance compared to related methods.

Our contributions are summarized as follows:

1) We propose a mutual information based learning objective and show that it achieves significantly better performance in terms of disentangling foreground and background.

2) We show that the synthesized image and segmentation mask pairs can be successfully used to train segmentation networks which achieve competitive performance compared to related methods.

2. Related Work

Layered GANs Generative adversarial networks (GANs) [18] have been proposed and improved towards highly photorealistic image synthesis [26, 7, 27] and explored for disentangling factors of variation towards controllable generation [17, 46, 26, 41, 23, 49, 48, 22], efficient representation learning [12], and saving of human annotations [55, 54]. To gain controllable generation at object level, layered GANs [52, 47, 6, 5] leverage the compositional nature of images and synthesize images by composing image layers, i.e. background and foreground objects. This compositional generation process is further developed to support 3D-aware scene generation [32, 37, 39, 38] and video generation [13]. Unlike most work that is interested in generation quality, our work focus on improving and employing 2D layered GANs towards learning segmentation with minimal human supervision.

In layered GANs, it is observed that compositional generation process is prone to degeneration to single layer generation [52, 10, 6]. This issue is alleviated by composing layers at feature level [13, 39], using weak supervision (e.g. bounding box annotation or background images) [32, 47, 5], or specific regularization (e.g. perturbation strategy as in PerturbGAN [6]). We propose a method based on mutual information maximization to mitigate this issue which can be successfully employed in layered GANs under the challenging unsupervised learning setting.

Object-Centric Scene Generation A rich body of work [16, 8, 19, 15, 34, 14, 3], grounded on variational autoencoders (VAEs) [30], simultaneously learn scene decomposition, object-centric representations, inter-object relationships, and multi-instance scene generation. Although these works also employ a compositional generation process and interested in the automatic segmentation, they are
mostly demonstrated on synthetic scenes but rarely evaluated on real-world datasets. In contrast, our work focuses on compositional generation and segmentation of real-world datasets.

**Unsupervised Segmentation** Earlier works [44, 21, 25, 2, 9] cast unsupervised object segmentation as coarsegmentation and utilise handcrafted features and shallow classifiers. With the advent of deep learning, per-pixel representation and clustering are simultaneously learned with deep neural networks [24, 40, 29] to achieve unsupervised segmentation. Apart from these works, other works learn segmentation from generative models. In particular, segmentation hint is extracted via cut and paste [42, 4, 6, 1], manipulating generation process [50, 36], erasing and redrawing [10], or maximizing inpainting error [45]. Our methods are based on generative models and learn segmentation from compositional generation process.

**3. Methods**

**3.1. Revisiting Layered GANs**

Let \( G : \mathcal{Z} \rightarrow \mathcal{X} \) denote a generative model that maps a latent variable \( z \in \mathcal{Z} \) to an image \( x \in \mathcal{X} \). In GANs, there is also a discriminator \( D : \mathcal{X} \rightarrow \mathbb{R} \) tasked to classify real and fake images. \( G \) and \( D \) are trained by playing the following adversarial game

\[
\min_G \max_D \mathbb{E}_{x \sim \text{plan}(x)} f(D(x)) + \mathbb{E}_{z} f(-D(G(z))),
\]

where \( f : \mathbb{R} \rightarrow \mathbb{R} \) is a concave function which differs for different kinds of GANs. In the end of adversarial learning, the images generated from \( G \) are high-fidelity and is hardly distinguished by \( D \).

To explicitly disentangle foreground objects from background, layered GANs decompose \( G \) into multiple generators responsible for generating layers which are further composed to synthesize full images. Formally, regarding two-layer generation, \( G \) is decomposed into background generator \( G_b \) and foreground generator \( G_f \) as \( G = (G_b, G_f) \). \( G_b \) generates background image \( x_b \) and \( G_f \) generates paired foreground image \( x_f \) and mask \( \pi \). The synthetic image \( x \) is completed by blending \( x_f \) and \( x_b \) with \( \pi \) as alpha map,

\[
x = C(x_b, x_f, \pi) = (1 - \pi) \odot x_b + \pi \odot x_f,
\]

where \( C \) denote the composition step. However, this naive layered GAN is prone to trivial decomposition such as “all as foreground” and “all as background” (Fig. 2 and Fig. 4), which requires further regularization as follows.

**Layer perturbation** Bielski & Favaro [6] propose to perturb the position of foreground object in composition step during training. Formally, this perturbed composition can be formulated as

\[
C(x_b, x_f, \pi; T) = (1 - T(\pi)) \odot x_b + T(\pi) \odot T(x_f),
\]

where \( T \) denotes a perturbation operator which PerturbGAN [6] instantiates as pixel-unit translation. In this paper, we employ restricted affine transformation. This method can effectively alleviate the all-as-foreground issue, because artefacts like sharp and unnatural border (Fig. 2(b)) emerges in the images under this circumstance and such inferior generation is penalized by discriminator. It should be noted that this perturbed composition does not hurt the fidelity of composed images in expected decomposition as long as the perturbation is imposed within a small range, because the relative position of foreground with respect to background can be varied up to a certain degree for most images. However, the perturbed composition is insufficient to address the all-as-background issue, which brings to our mutual information maximization as in the next section.

**3.2. Improving Layered GANs with Mutual Information Maximization**

Fig. 3 presents the framework of our model. Generally, we improve layered GANs with in terms of disentangling foreground object and background from a special design of generative model and mutual information maximization, which is described as follows.

Our generative model takes as input two kinds of latent code, denoted as \( z \) and \( c \), respectively. In particular, \( z \) is shared across \( G_b \) and \( G_f \), whereas \( c \) is private to \( G_f \). We therefore refer to \( z \) and \( c \) as public code and private code, respectively (Fig. 3). Based on this design, we introduce an additional learning objective aiming at maximizing mutual information between the composed image \( x \) and the private code \( c \) and mutual information between the generated mask \( \pi \) and the private code \( c \),

\[
\max I(x, c) + I(\pi, c).
\]

This learning objective can penalize \( G_f \) from always generating zero mask, which is shown as following.

For the first term in Equ. (4), it is the composed image \( x \) that is measured the mutual information with \( c \). If \( G_f \) produces all-zero mask, i.e. \( \pi = 0 \), the composed image would only contain background, i.e. \( x = x_b \), and \( I(x, c) = I(x_b, c) = 0 \) since \( c \) is private to foreground generation process. Therefore, non-zero mask is essential for maximizing \( I(x, c) \). For the second term, the mutual information can be rewritten as difference between two entropies, \( I(\pi, c) = H(\pi) - H(\pi | c) \). Note that our generative model is deterministic, meaning there is no stochastic in \( \pi \) given \( c \), i.e. \( H(\pi | c) = 0 \). Hence, maximizing \( I(\pi, c) \) is equivalent to maximizing \( H(\pi) \). If \( G_f \) constantly generates all-zero masks, \( H(\pi) \), which depicts the diversity of
Figure 3. **Framework.** Our generative model comprises background generator $G_b$, foreground generator $G_f$, and composer with perturbation $C$. The public code $z$ is shared across foreground and background generation, while the private code $c$ is only used for foreground generation. Our generative model is mainly trained via adversarial learning against a discriminator $D$, denoted as $\mathcal{L}_{adv}$, and regularized by mutual information loss, denoted as $\mathcal{L}_{mi,\pi}$ and $\mathcal{L}_{mi,x}$, whose estimation is assisted by $E_{\pi}$ and $E_x$. $E_x$, which is omitted in the figure, shares backbone parameters with $D$ and bifurcates at the top layers. In the alternate training, background generation is regularized by segmentation network $F$, denoted as $\mathcal{L}_{bg}$.

In practice, computing the mutual information, $I(x,c)$ and $I(\pi,c)$, is intractable. Following InfoGAN [11], we instead maximize the variational lower bound of the mutual information, which is realized as following. First, neural networks $E_{\pi}$ and $E_x$ are introduced to approximate the posterior distribution $p(c|\pi)$ and $p(c|x)$, respectively. Second, $G$ and $E_{\pi}$ and $E_x$ are learned by minimizing the following loss

$$L_{mi} = \mathbb{E}_{z,c} \log E_x(c|x) + \log E_{\pi}(c|\pi) \quad (5)$$

where $E_x(c|x)$ and $E_{\pi}(c|\pi)$ denote the likelihood of a sample $c$ under the approximate posterior distribution of $p(c|\pi)$ and $p(c|x)$, respectively. At the implementation level, $E_x$ shares backbone network parameters with discriminator $D$ and bifurcates at the last two layers, as commonly done in [11, 33, 47, 5]. More implementation details can be found in the appendices.

In addition to penalizing trivial masks, introducing private code and maximizing $I(x,c)$ is also favorable for foreground-oriented clustering. In particular, $c$ can be designed as discrete variable and its prior distribution can be a uniform categorical distribution. In this way, $c$ is interpreted as clustering category and $E_x$ is regarded as a clustering network trained from synthetic data. Note that the unique design of our generative model, where $c$ is only used for foreground generation, makes the clustering based only on the variety of foreground objects and not interfered by backgrounds. This mechanism conforms to the logics of category annotation for major datasets, where it is foreground objects that are categorized rather than background. Experimental results in Section 4.3 (Table 4) justifies the merit of our foreground-oriented clustering.

### 3.3. Alternate Training of Layered GAN and Segmentation Network

**Training a layered GAN** The improved Layered GAN is optimized by updating generator and discriminator alternately,

$$\min_G \mathcal{L}_{adv}^{(G)} + \gamma_{mi}\mathcal{L}_{mi}^{(G)} + \gamma_b\mathcal{L}_b^{(G)}, \quad \min_{D,E_{\pi},E_x} \mathcal{L}_{adv}^{(D)} + \gamma_{mi}\mathcal{L}_{mi}^{(D)} + \gamma_b\mathcal{L}_b^{(D)}, \quad (6)$$

where $\mathcal{L}_{mi}$ denotes the mutual information loss as defined in Eqn. (5), $\mathcal{L}_{adv}$ denotes the adversarial loss for which hinge loss is employed, i.e. $\mathcal{L}_{adv}^{(G)} = -\mathbb{E}_{z,c} D(G(z,c))$, and $\mathcal{L}_{adv}^{(D)} = \mathbb{E}_{x \sim p_{data}(x)} \max(0, 1 - D(x)) + \mathbb{E}_x \max(0, 1 + D(G(z,c)))$, $\mathcal{L}_b^{(G)}$ denotes the binarization loss $\mathcal{L}_b^{(G)} = \mathbb{E}_{z,c} \min(\pi, 1 - \pi)$ to encourage the binarization of masks as in [6], and $\gamma_{mi}$, $\gamma_b$ denote the corresponding loss weights.

**Training a Segmentation Network** Improved layered GAN can synthesize paired images and segmentation masks by sampling latent variables from prior distribution and forwarding them to the generator. This procedure is repeated for $N$ times to generate a synthetic dataset $\mathcal{D}_{syn} = \{(x_i, \pi_i)\}_{i=1}^N$ which is further used for training a foreground-background segmentation network $F : \mathcal{X} \rightarrow \mathcal{Y}$. 

2527
Algorithm 1: Alternate training

1. for $k = 1, \ldots, K$ do
2.   Set $\hat{\gamma}_b = 0$ if $k = 1$ else $\hat{\gamma}_b = \gamma_b$
3.   for $n = 1, \ldots, N_{gan}$ do
4.     Update $G_b$ and $G_f$ with loss
5.     \begin{equation}
L^{(G)} = L^{(G)}_{adv} + \gamma_m L^{(G)}_{mi} + \gamma_b L^{(G)}_{bg}
\end{equation}
6.     Update $D, E_x$ and $E_\pi$ with loss
7.     \begin{equation}
L^{(D)} = L^{(D)}_{adv} + \gamma_m L^{(E_x,E_\pi)}
\end{equation}
8.   for $n = 1, \ldots, N_{seg}$ do
9.     Generate a minibatch of synthetic dataset $B$
10. Update $F$ on $B$ with loss
11. \begin{equation}
L^{(F)} = E_{(x,\pi) \sim B} [(1 - \pi) \log(1 - F(x)) + \pi \log(F(x))]
\end{equation}

$[0, 1]^{H \times W}$ as follows,
\begin{equation}
\min_F E_{(x,\pi) \sim D_{syn}} [(1 - \pi) \log(1 - F(x)) + \pi \log(F(x))].
\end{equation}
Since the layered GAN approximates the real data distribution and produces reasonable masks, the optimized $F$ could fairly generalize to real images.

Alternate Training As presented in Algorithm 1, an alternate training schedule is employed to exploit the synergy between layered GAN and segmentation segmentation network. In particular, the training is conducted for $K$ rounds. In the initial round, a layered GAN is trained without the help of segmentation network. The initially learned layered GAN is sufficient to generate synthetic dataset for training a segmentation network. In the rest rounds, the segmentation network is used to regularize the background generation with an additional loss term
\begin{equation}
L^{(G)}_{bg} = E_{z,e} \log(1 - F(x_b)),
\end{equation}
where the parameters of segmentation network are fixed. This loss encourages the generated background pixels, $x_b$, to be classified as background and helps to stabilize the learned disentanglement of foreground and background. The training of layered GAN is then followed by another time of updating segmentation network. In this way, layered GAN and segmentation network are alternately trained. More training details are available in the appendices.

4. Experiments

4.1. Settings

Datasets Our method is evaluated on a variety of datasets including Caltech-UCSD Birds 200-2011 (CUB) [51], Stanford Cars [31], Stanford Dogs [28], and Amazon Picking Challenge (APC) [53]. CUB, Stanford Cars and Stanford Dogs are fine-grained category datasets for bird, car, and dog, respectively. Most images in these three datasets only contain single object. APC is a image dataset collected from robot context which contains images of single object either on a shelf or in a tray. Unlike the above four datasets, objects in APC belong to a variety of categories such as ball, scissors, book, etc. Ground truth segmentation masks of CUB and APC are available from their public release, whereas pre-trained Mask R-CNN are employed to approximate the ground truth segmentation of Stanford Cars, and Stanford Dogs, as in previous work [47, 5]. More details about the datasets are available in the appendices.

Evaluation metrics Our methods are quantitatively evaluated with respect to generation quality and segmentation performance. To measure the quality of generation, Fréchet Inception Distance (FID) [20] between 10k synthetic images and all the images are computed. The segmentation performance is evaluated with per-pixel accuracy (ACC), foreground intersection over union (IoU) between predicted masks and ground truth ones, and mean IoU (mIoU) across foreground and background. Moreover, adjusted random index (ARI) [19] and mean segmentation covering (MSC) [15] are computed to compare the segmentation performance level of our method to that of VAE-based methods. In addition to above metrics, normalized mutual information (NMI) is computed to measure the category clustering performance of our method which is compared to other layered GANs.

4.2. Ablation Studies

We conduct multiple ablation studies to investigate (i) the effect of mutual information maximization, (ii) the influence of different choices for private code $c$ distribution, and (iii) the effect of alternate training by controlled experiments which are mainly conducted in CUB dataset at $64 \times 64$ resolution.

Table 1. Ablation study with respect to mutual information maximization on CUB. $\gamma_{mi}$ denotes the loss weight of mutual information maximization. $L_{size}$ denotes the mask size loss.

<table>
<thead>
<tr>
<th>$L_{size}$</th>
<th>0.1</th>
<th>0.5</th>
<th>1.0</th>
<th>5.0</th>
<th>10.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>FID ↓</td>
<td>19.4</td>
<td>19.1</td>
<td>17.6</td>
<td>11.6</td>
<td>15.5</td>
</tr>
<tr>
<td>bg-FID ↑</td>
<td>20.9</td>
<td>18.6</td>
<td>17.4</td>
<td>99.5</td>
<td>107.1</td>
</tr>
<tr>
<td>IoU ↑</td>
<td>6.3</td>
<td>0.1</td>
<td>68.6</td>
<td>66.5</td>
<td>66.3</td>
</tr>
</tbody>
</table>

Challenge (APC) [53]. CUB, Stanford Cars and Stanford Dogs are fine-grained category datasets for bird, car, and dog, respectively. Most images in these three datasets only contain single object. APC is a image dataset collected from robot context which contains images of single object either on a shelf or in a tray. Unlike the above four datasets, objects in APC belong to a variety of categories such as ball, scissors, book, etc. Ground truth segmentation masks of CUB and APC are available from their public release, whereas pre-trained Mask R-CNN are employed to approximate the ground truth segmentation of Stanford Cars, and Stanford Dogs, as in previous work [47, 5]. More details about the datasets are available in the appendices.

Evaluation metrics Our methods are quantitatively evaluated with respect to generation quality and segmentation performance. To measure the quality of generation, Fréchet Inception Distance (FID) [20] between 10k synthetic images and all the images are computed. The segmentation performance is evaluated with per-pixel accuracy (ACC), foreground intersection over union (IoU) between predicted masks and ground truth ones, and mean IoU (mIoU) across foreground and background. Moreover, adjusted random index (ARI) [19] and mean segmentation covering (MSC) [15] are computed to compare the segmentation performance level of our method to that of VAE-based methods. In addition to above metrics, normalized mutual information (NMI) is computed to measure the category clustering performance of our method which is compared to other layered GANs.

4.2. Ablation Studies

We conduct multiple ablation studies to investigate (i) the effect of mutual information maximization, (ii) the influence of different choices for private code $c$ distribution, and (iii) the effect of alternate training by controlled experiments which are mainly conducted in CUB dataset at $64 \times 64$ resolution.

The effect of mutual information maximization Table 1 and Fig. 4 presents the results of ablation study with respect to mutual information maximization, where $\gamma_{mi}$ is varied or mask size loss $L_{size}$ is used to replace mutual information
maximization. In addition to FID and IoU, we also compute bg-FID which is the FID between synthetic background image and realistic images. To some extent, a bg-FID that is close to FID can manifest that “all as background” occurs. Therefore, we would expect there is sufficient gap between bg-FID and FID. It can be seen from Fig. 4 that “all as background” still occurs when using mask size loss. This is also evidenced by the poor segmentation performance and the negligible gap between FID and bg-FID (Table 1). In contrast, the utilization of mutual information loss with appropriate weight (e.g. \( \gamma_{mi} = 1, 5, 10 \)) can successfully suppress the occurrence of “all as background”. It is noteworthy that the benefits from mutual information loss saturates and the FID score, which reflects the fidelity and diversity of synthetic images, drops as \( \gamma_{mi} \) increases beyond 1.0. We attribute this to adversarial learning being overwhelmed by mutual information maximization. To trade off between the quality of synthetic fidelity and suppression of degeneration, a proper \( \gamma_{mi} \) needs to be chosen, but the secure range of \( \gamma_{mi} \) is quite wide as the performances of \( \gamma_{mi} \) from 1 to 10 are quite close. Additional results on other datasets are available in the supplementary material.

### Prior distribution of \( c \)

We investigate the influence of different choices for prior distribution of \( c \). In particular, normal distribution and categorical distribution are compared and the dimension (or number of categories) of \( c \) is varied. Results in Table 2 show that our method achieves similar performance with different settings, suggesting that the proposed mutual information maximization is robust to different choices for prior distribution of \( c \). Although normal distribution leads to marginally higher performance than categorical distribution, the categorical distribution advocates the foreground-oriented clustering as discussed in 3.2 and makes our method comparable to other layered GANs [47, 5].

#### The effect of alternate training

We compare the performance of our methods with and without the alternate training strategy on CUB, Stanford Dogs, and Stanford Cars at 64 × 64 resolution. Table 3 shows that the alternate training helps to improve the segmentation performance without compromising the generation quality.

### 4.3. Results in Fine-Grained Category Datasets

Results of our method in three fine-grained category datasets, CUB, Stanford dogs, and Stanford Cars, are presented and compared to that of other layered GANs, FineGAN [47] and OneGAN [5]. All of the experiments are conducted at 128 × 128 resolution. It should be noted that FineGAN and OneGAN requires all of or a portion of training images annotated with bounding boxes, which should be regarded as weak supervision. On the contrary, our models only learn from images, which is should be thought of as unsupervised learning. Table 4 compares our method to FineGAN and OneGAN in the aspect of generation quality, segmentation performance, and clustering performance with FID, IoU, and NMI, respectively, and Fig. 5 presents the qualitative generation results of our method.

Despite in the more challenging unsupervised learning
setting, our method is still able to learn disentangled foreground and background and synthesize quality paired images and segmentation masks (Fig. 5). Our method even achieves higher generation quality than FineGAN and OneGAN in CUB and Stanford Cars, with 7.6 and 5.2 higher FID, respectively. The segmentation performance of our method is still competitive to that of FineGAN and OneGAN. It is notable that our method even outperforms OneGAN in CUB with large margin (14.2 IoU) and achieves close performance to OneGAN on Stanford Dogs and Stanford Cars. We also take the trained $E_s$ as a clustering network and evaluate its category clustering performance (Table 4). It shows that our method constantly obtains higher NMI score than FineGAN and OneGAN. This result justifies the merit of our foreground-oriented clustering as the ground truth category label is annotated based on foreground rather than background as discussed in Section 3.2.

4.4. Object Segmentation Results

Our methods are compared to other unsupervised segmentation methods on CUB at 128 × 128 resolution. Quantitative and qualitative results are presented in Table 5 and Fig. 6, respectively. It is noteworthy that the method proposed by Vojnov et al. [50] requires manual inspection which should be regarded as external supervision. Our method outperform PerturbGAN [6] with a large margin, showing the merit of our proposed mutual information maximization as discussed in the ablation study. It is notable that performance level of our method is comparable to the state-of-the-art method [45, 36]. In particular, our method gain significantly higher performance than Melas-Kyriazi et al. [36] with 3.3 IoU on CUB.

4.5. Comparison to VAE-based methods

Our methods are evaluated and compared to VAE-based models including MONet [8], GENESIS [15], SLOT ATTENTION [35], and GENESIS-V2 [14] in APC dataset. The quantitative and qualitative results are presented in Table 6, Fig. 7, and Fig. 8. The majority of VAE-based models perform poorly in realistic dataset as suggested in Table 6. GENESIS-V2 is the latest work that is successfully demonstrated in real-world dataset like APC. Therefore, our...
methods are mainly compared to GENESIS-V2. To obtain the results of GENESIS-V2, we use its released codes and models\(^1\).

\(^1\)https://github.com/applied-ai-lab/genesis

It can be seen from Fig. 7 and Fig. 8 that our methods are successfully applied to APC dataset which contains multi-category objects. These results show that our methods are not limited to single-category dataset. It is not surprising that our methods, as based on GAN, generate synthetic images of significant higher quality than GENESIS-V2 does, as suggested by the more visual pleasing images in Fig. 7 and much lower FID score in Table 6. Furthermore, our methods achieve higher segmentation performance than GENESIS-V2 and predict finer segmentation results than GENESIS-V2 does, as shown in Fig. 8.

4.6. Limitations

Our generative model is limited to synthesize single object and single instance images. Our method also inherits the characteristic problems of GAN training including instability and requiring certain number of unlabelled training images. In the future, we plan to extend our model into multi-instance image generation by modelling inter-object relationship and also improve the stability of training process with powerful architectures and techniques.

5. Conclusion

In this paper, we propose a method to improve the capability of layered GANs to disentangle foreground and background. This improved layered GAN can successfully generate synthetic datasets to train segmentation networks which can be, in return, employed to further stabilize the training of layered GANs. Our method achieves competitive generation quality and segmentation performance compared to related methods on a variety of single-object image datasets. We plan to extend our method to handle multiple object instances and the noisy data in future work.

Acknowledgement This work was supported by the National Key R&D Program of China under Grant 2018AAA0102801, National Natural Science Foundation of China under Grant 61620160005, and EPSRC Visual AI grant EP/T028572/1.
References


