Enhancing TripleGAN for Semi-Supervised Conditional Instance Synthesis and Classification

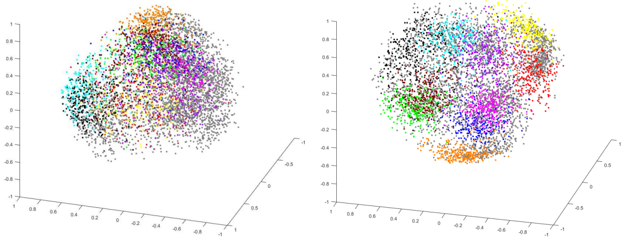
Si Wu[12] Guangchang Deng[1] Jichang Li[1] Rui Li[2] Zhiwen Yu[1] Hau-San Wong[2]
[1]School of Computer Science and Engineering, South China University of Technology
[2]Department of Computer Science, City University of Hong Kong
cswusi@scut.edu.cn, csgc@mail.scut.edu.cn, cslijichang@mail.scut.edu.cn
ruili52-c@my.cityu.edu.hk, zhwyu@scut.edu.cn, cshswong@cityu.edu.hk

Abstract

Learning class-conditional data distributions is crucial for Generative Adversarial Networks (GAN) in semi-supervised learning. To improve both instance synthesis and classification in this setting, we propose an enhanced TripleGAN (EnhancedTGAN) model in this work. We follow the adversarial training scheme of the original TripleGAN, but completely re-design the training targets of the generator and classifier. Specifically, we adopt feature-semantics matching to enhance the generator in learning class-conditional distributions from both the aspects of statistics in the latent space and semantics consistency with respect to the generator and classifier. Since a limited amount of labeled data is not sufficient to determine satisfactory decision boundaries, we include two classifiers, and incorporate collaborative learning into our model to provide better guidance for generator training. The synthesized high-fidelity data can in turn be used for improving classifier training. In the experiments, the superior performance of our approach on multiple benchmark datasets demonstrates the effectiveness of the mutual reinforcement between the generator and classifiers in facilitating semi-supervised instance synthesis and classification.

1. Introduction

Significant advances in deep learning techniques have resulted in its wide adoption in a broad set of applications, most notably in computer vision [14] [31] [34] and natural language processing [42]. However, in view of the need to collect a massive amount of labeled data for most fully-supervised deep learning models, semi-supervised learning represents an effective approach to avoid the need for extensive manual annotations. This is due to the capability of semi-supervised learning to capture the characteristics of a dataset through a small number of labeled instances, together with a large set of unlabeled instance. A number of previous methods have been developed to learn discriminative representations, explore the underlying manifold structures, and infer the labels of the unlabeled data, such as [41] [12] [18] [1] [39]. However, the quality of the unlabeled data will have a significant effect on the performance of semi-supervised learning, and incorporation of low-quality data in the training process could lead to ambiguous or even incorrect decisions.



(a) TripleGAN*

(b) EnhancedTGAN

Figure 1. The embedding of the unlabeled training data and synthesized data on CIFAR-10 with 4000 labels. The features of the last hidden layer of the classifier network are projected to 3D using PCA. The unlabeled samples are marked gray, and different colors denote different classes of synthesized samples. We implement TripleGAN [17] in our configuration environment as the baseline which is referred to as TripleGAN*. One can observe that the proposed EnhancedTGAN performs better than TripleGAN* in learning class-conditional distribution, since the synthesized data can match the unlabeled data in sub-figure (b).

Recent applications of generative adversarial network (GAN) [8] [29] [43] [21] [22] to semi-supervised learning have shown promise, due to the capability of GAN to synthesize high-quality samples by learning the probability distribution of a data set. To perform semi-supervised data synthesis, Odena [27] modified the discriminator network to classify the synthesized data to the $K+1$-th class, while in [35] the predicted class probability distribution for the syn-

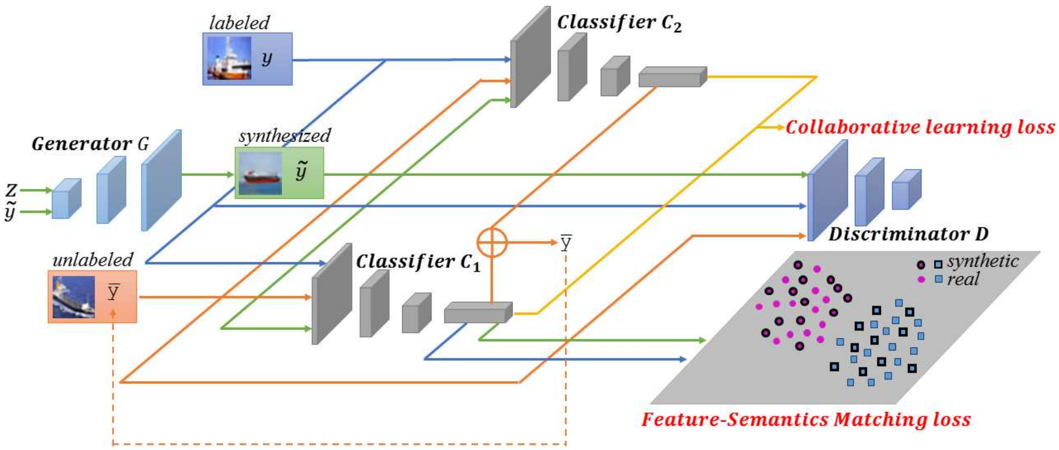


Figure 2. Illustration of the structure of the proposed EnhancedTGAN model for semi-supervised conditional instance synthesis and classification.

thesized data is forced to be uniform. To avoid the case in which the discriminator also needs to predict the class label for real data, Li et al. [17] proposed the TripleGAN model, in which a classifier was incorporated into the adversarial training process. However, a limited number of labeled instances are not sufficient for learning the class-conditional probability distribution of the categories. In this case, there may exist a domain shift in which the divergence between the real and synthesized data distributions is significant. In particular, the domain discrepancy may be significant in the early phase of the training process, and the generator and classifier may negatively affect each other. As a result, it is important, for the purpose of facilitating semi-supervised synthesis and classification, to effectively match the statistics of real and synthesized data, and this work presents a feature-semantics matching approach to achieve this objective as shown in Figure 1.

In this work, we propose an enhanced TripleGAN (EnhancedTGAN) model for improving semi-supervised conditional instance synthesis and classification. We follow the adversarial training scheme of the TripleGAN model in general, but re-design the overall loss functions of the generator and classifiers. Our feature-semantics matching approach is able to reduce the risk of mode collapse and improve the synthesis of the instances of each class. Specifically, we adopt class-wise mean feature matching to regularize the generator, such that the class-conditional distribution of the synthesized data can match with that of the real data for each class in the latent space learnt by a classifier, instead of the discriminator. In addition, we further include a semantic matching term to ensure the semantics consistency of the synthesized data between the generator and the classifier, which is a prerequisite for improving classifier training. On the other hand, a better classification model can provide more accurate categorical information on a large number of unlabeled instances, which leads to better guidance for the generator. For this purpose, we include two classifiers in our model which operate in a collaborative learning fashion. The classifiers can learn from each other by penalizing the divergence between the predicted class probability distributions, and the consensus predictions on the unlabeled data are more accurate than individual predictions in most cases. As a result, the generator and the classifiers can mutually reinforce each other. The structure of our model is illustrated in Figure 2. Our experiments verify the effectiveness and superiority of the proposed model. The main contributions of this work are summarized as follows:

- We propose a feature-semantics matching approach through which the generator can more effectively learn the class-conditional data distributions.
- In addition to the synthesized high-fidelity instances for training data augmentation, collaborative learning between the classifiers can also lead to more accurate classification on the unlabeled data, which in turn provides better guidance for the generator.
- The proposed enhanced TripleGAN model improves the state-of-the-art results in both semi-supervised instance synthesis and classification on multiple widely used benchmarks.

2. Related Work

Recently, various strategies have been applied to improve semi-supervised deep learning. Unsupervised learning can be used as an auxiliary task for exploring the structure of a dataset, and thus the generalization capability of the classification model can be improved. To formulate the unsupervised learning loss function for unlabeled samples, Rasmus et al. [30] proposed the Γ-model, in which a consistency regularization term was adopted to penalize the inconsistent predictions of the Ladder network [37] in inputs with and without noise. Similar to the Γ-model, Laine and Aila [15] proposed the Π-model, which regularized the model outputs of the training samples under different dropout and augmentation conditions. To provide more stable training targets for the unsupervised loss, Laine and Aila further proposed the Temporal-Ensembling model. In this model, self-ensembling of the network was applied to complement supervision. The ensemble of predictions at different epochs are expected to be more accurate, and can thus be used as the training targets for unlabeled samples. In contrast to the exponential moving average of the network predictions on each sample, Tarvainen and Valpola [36] proposed the Mean-Teacher model to average the network weights, and the resulting model can be considered as a teacher to provide the training targets for the unlabeled samples. Based on the assumption that similar samples should be located in the same cluster in the latent space, Luo et al. [19] proposed the Smooth Neighbors on Teacher Graphs (SNTG) method to build a graph based on the predictions of a teacher network for measuring the similarity between unlabeled data points. To ensure smoothness on the data manifold, the contrastive loss was used to ensure that neighbors had consistent predictions, while non-neighbors were pushed away from each other. In addition, it is common to incorporate the locally-Lipschitz condition through penalizing inconsistent predictions of unlabeled samples with different perturbations. Miyato et al. [24] [23] proposed the Virtual Adversarial Training (VAT) based regularization method to improve the local smoothness of the predicted class probability distribution by applying perturbation in the adversarial direction with respect to the classification model. VAT can be incorporated into existing semi-supervised learning networks, and yields impressive results. On the other hand, Park et al. [28] developed the Virtual Adversarial Dropout (VAdD) approach to reconfigure the neural network and minimize the divergence between the obtained network and the original network for increasing the sparsity of the overall network.

Deep generative models have been recently applied to semi-supervised learning. In [10], Kingma et al. adopted the Variational Auto-Encoder (VAE) model [11] to treat class label as an additional latent variable in the process of learning the generative model. In [27], Odena modified the discriminator network to simultaneously distinguish real samples from synthesized samples and predict the corresponding class labels. Springenberg [35] proposed the Categorical Generative Adversarial Network (CatGAN) to make the discriminator assign high-confidence class labels for the real samples, while forcing the predicted class probability distributions on the synthesized samples to be uniform. Salimans et al. [32] proposed a variety of training techniques to improve the GAN training procedure, which lead to improvement in semi-supervised learning and sample synthesis. Furthermore, Wei et al. [38] improved the Wasserstein GAN [2] by including a consistency term with respect to the discriminator responses for enforcing Lipschitz continuity. To prevent the discriminator from playing two roles of identifying synthesized samples and predicting class labels for real samples in a minimax game, Li et al. [17] incorporated a classifier as an additional player into the game, and proposed the Triple Generative Adversarial Net (TripleGAN). Dumoulin et al. [5] proposed the Adversarially Learned Inference (ALI) model, in which a generation network learnt the mapping from the latent space to the data space, while an inference network learnt the inverse mapping. These two networks were jointly optimized with a discriminative network in an adversarial process. To learn the joint distribution between samples and labels, Gan et al. [7] proposed the Triangle Generative Adversarial Network (TriangleGAN), in which two generators were adopted to learn the conditional distributions between samples and labels, and two discriminators were used to identify the types of fake pairs between real (fake) samples and fake (real) labels. Instead of matching the real and fake data distributions, Dai et al. [4] proposed a complementary generator which was trained by minimizing the KL divergence between the distributions, such that the generated samples were located in the low-density region in the latent space, and the diversity of the training data was increased.

TripleGAN is the most related work to our proposed approach. However there are significant differences between them. Although we follow the adversarial training scheme of TripleGAN in general, we completely redesign the overall loss function of the generator by including feature-semantics matching for effective and efficient learning of the class-conditional data distributions. In addition, we include two classifiers which collaboratively learn from each other to provide more accurate categorical information for the generator. As a result, the generator and classifier mutually reinforce each other to facilitate semi-supervised instance synthesis and classification.

3. Method

Inspired by the method in [6], the maximum mean discrepancy measure was used for training GANs. Feature matching has shown effectiveness in addressing the instability problem in GAN. The objective function defined below can be applied to force the generator G to synthesize data that matches the statistics of the real data [32] [3]:

$$\left\| E_{x \sim p_{data}} f_D(x) - E_{z \sim p_z} f_D(G(z)) \right\|, \tag{1}$$

where p_{data} denotes the distribution of real data x, p_z denotes the distribution of random vector z, e.g., $U[0,1]$, $G(z)$ denotes a synthesized sample from z, and $f_D(\cdot)$ denotes the features associated with the hidden layer of the discriminator D. The center of synthesized data points is forced to match that of real data points in the latent space learnt by the discriminator. However, there are two main issues when applying the above formulation to our task. On one hand, the categories of the instances are not taken into account during the process of matching the marginal distributions. On the other hand, in addition to class-conditional instance synthesis, another objective is to perform accurate classification on the unlabeled data, while feature matching in the space learnt by the discriminator cannot directly improve classification. In this section, we introduce the EnhancedTGAN model to improve both semi-supervised conditional instance synthesis and classification.

3.1. Feature-Semantics Matching

In our setting, only a small portion of the training samples are labeled. Let $\mathrm{x} \sim p_u$ denote unlabeled samples, and $(x, y) \sim p_l$ denote the labeled data pair, where y denotes the label of sample x. Our EnhancedTGAN consists of the following four modules: the generator G, discriminator D, and classifiers C_1 and C_2. We slightly modify the adversarial training scheme of the TripleGAN model. Specifically, the generator G synthesizes new instances by sampling pairs of random vector and class label $(z, \tilde{y})$ from a pre-specified distribution p_g. The two classifiers C_1 and C_2 collaboratively learn from each other, and produce the consensus prediction $\bar{\mathrm{y}}$ of the input data.

To improve class-conditional instance synthesis, we optimize the generator by including the class-wise mean feature matching term defined as follows:

$$\begin{aligned}\ell_{feaMat}(\theta_G) = \sum_k \Big\| &\mathbb{E}_{(x,y)\sim p_l}\big[\mathbf{1}(y,k) f_{C_1}(x)\big] \\ &- \mathbb{E}_{(z,\tilde{y})\sim p_g}\big[\mathbf{1}(\tilde{y},k) f_{C_1}(G(z,\tilde{y}))\big]\Big\|,\end{aligned} \tag{2}$$

where k denotes the class index, $f_{C_1}(\cdot)$ denotes the features on the hidden layer of the classifier C_1, and the function $\mathbf{1}(\cdot,\cdot)$ returns 1 if the inputs are equal and 0 otherwise. S-ince the number of labeled instances is small, we can use moving historical averages to obtain more stable means for them. The main advantage of ℓ_{feaMat} is to avoid the mode collapse problem where the generator always outputs the same point. Another advantage is to increase the separability of different classes of synthesized data. In order to utilize the synthesized samples for training the classifiers, their semantics from the perspectives of the classifier and generator should be consistent. To enforce this consistency, we adopt a semantics matching term to regularize the generator as follows:

$$\ell_{semMat}(\theta_G) = \mathbb{E}_{(z,\tilde{y})\sim p_g}\big[-\tilde{y}\log \bar{p}_C(G(z,\tilde{y}))\big], \tag{3}$$

where

$$\bar{p}_C(x) = \texttt{avg-pool}\big(p_{C_1}(x), p_{C_2}(x)\big), \tag{4}$$

and $p_{C_1}(\cdot)$ ($p_{C_2}(\cdot)$) denotes the predicted class probability distribution by classifier C_1 (C_2). The average pooling of the classifier predictions can be expected to be more accurate in most cases. After including the adversarial training term with the discriminator, the optimization of the generator can be formulated as follows:

$$\begin{aligned}\min_G \ \frac{1}{2}\mathbb{E}_{(z,\tilde{y})\sim p_g}\big[\log(1 - D(G(z,\tilde{y}),\tilde{y}))\big] \\ + \eta\ell_{feaMat} + \nu\ell_{semMat},\end{aligned} \tag{5}$$

where the weighting factors η and ν are used for controlling the relative importance of the corresponding terms.

3.2. Collaborative Learning of Classifiers

Different from the TripleGAN model, we include two classifiers in our model, due to the reason that they can provide the training targets of the unlabeled instances for each other via collaborative learning. Existing works have demonstrated that collaborative learning is capable of facilitating semi-supervised classification.

Similar to the given labeled instances, the synthesized instances can also be utilized because of the known labels. The classifier can be enhanced by including the instances from the generator. The loss measure for supervised learning is the cross entropy between the given labels and the predicted distribution. Furthermore, the model can also learn from the unlabeled samples by minimizing the conditional entropy with respect to the posterior probability distribution. Therefore, we define the term for classification evaluation as follows:

$$\begin{aligned}\ell_{classify}(\theta_{C_1}) =& \mathbb{E}_{(x,y)\sim p_l}\big[-y\log p_{C_1}(x)\big] \\ &+ \mathbb{E}_{(z,\tilde{y})\sim p_g}\big[-\tilde{y}\log p_{C_1}(G(z,\tilde{y}))\big] \\ &+ \mathbb{E}_{\mathrm{x}\sim p_u}\big[-p_{C_1}(\mathrm{x})\log p_{C_1}(\mathrm{x})\big].\end{aligned} \tag{6}$$

The classifiers tend to be confident on the unlabeled samples. To stabilize the estimation of the conditional entropy, a smoothness regularization term ℓ_{smoReg} is defined as follows:

$$\ell_{smoReg}(\theta_{C_1}) = \mathbb{E}_{\mathrm{x}\sim p_u}\left[\max_{\|\gamma\|\le\xi} \mathtt{KL}\big(p_{C_1}(\mathrm{x}) \| p_{C_1}(\mathrm{x}+\gamma)\big)\right], \tag{7}$$

where the constant ξ is used to control the intensity of the adversarial perturbation γ, and $\mathtt{KL}(\cdot\|\cdot)$ denotes the Kullback-Leibler (KL) divergence. Similar to [24], the perturbation is generated in the direction most sensitive to the classifier prediction, and the KL divergence is used to measure the prediction difference with respect to the classifier for cases with and without perturbation. As a result, the output of the classifier will become smooth in the neighborhood of unlabeled samples.

Algorithm 1 The proposed EnhancedTGAN model for semi-supervised conditional instance synthesis and classification.

```
Input: Labeled data $X_l$ and unlabeled data $X_u$.
Initialize: Generator $G$, discriminator $D$, classifiers $C_1$ and $C_2$, learning rate $\zeta_G$, $\zeta_D$ and $\zeta_C$, and batch size $b_l$, $b_u$ and $b_g$ for labeled, unlabeled and synthesized samples, respectively.
for $n = 1$ to $N$ do
    Sample labeled instances $\{(x, y)\}$ of size $b_l$ from $X_l$, unlabeled instances $\{\mathrm{x}\}$ of size $b_u$ from $X_u$, and random vectors $\{(z, \tilde{y})\}$ of size $b_g$ from the uniform distribution.
    for each mini-batch $B$ do
        Evaluate classifier predictions $p_{\theta_{C_1}}$ and $p_{\theta_{C_2}}$ for $x$, $\mathrm{x}$ and $G(z, \tilde{y})$.
        Compute the consensus results $\bar{p}_C$ and the corresponding one-hot label $\bar{\mathrm{y}}$ for $\mathrm{x}$.
        Update the discriminator $D$ by using Adam [9]
            $\theta_D \leftarrow Adam\Big(\nabla_{\theta_D}\Big(\sum_{(x,y)} \log D(x,y) + \frac{1}{2}\sum_{(z,\tilde{y})} \log(1 - D(G(z,\tilde{y}),\tilde{y})) + \frac{1}{2}\sum_{(\mathrm{x},\bar{\mathrm{y}})} \log(1 - D(\mathrm{x},\bar{\mathrm{y}}))\Big), \theta_D, \zeta_D\Big).$
        Update the classifiers $C_1$ and $C_2$ by using Adam
            $\theta_{C_1} \leftarrow Adam\Big(\nabla_{\theta_{C_1}}\Big(\frac{1}{2}\sum_{(\mathrm{x},\bar{\mathrm{y}})} \bar{p}_C(\mathrm{x}) \log(1 - D(\mathrm{x},\bar{\mathrm{y}})) + \ell_{classify} + \lambda\ell_{smoReg} + \mu\ell_{conReg}\Big), \theta_{C_1}, \zeta_C\Big),$
            $\theta_{C_2} \leftarrow Adam\Big(\nabla_{\theta_{C_2}}\Big(\frac{1}{2}\sum_{(\mathrm{x},\bar{\mathrm{y}})} \bar{p}_C(\mathrm{x}) \log(1 - D(\mathrm{x},\bar{\mathrm{y}})) + \ell_{classify} + \lambda\ell_{smoReg} + \mu\ell_{conReg}\Big), \theta_{C_2}, \zeta_C\Big).$
        Update the generator $G$ by using Adam
            $\theta_G \leftarrow Adam\Big(\nabla_{\theta_G}\Big(\frac{1}{2}\sum_{(z,\tilde{y})} \log(1 - D(G(z,\tilde{y}),\tilde{y})) + \eta\ell_{feaMat} + \nu\ell_{semMat}\Big), \theta_G, \zeta_G\Big).$
    end for
end for
Return $\theta_G$, $\theta_D$, $\theta_{C_1}$ and $\theta_{C_2}$.
```

To encourage the classifiers to learn from each other, we further define a consistency regularization term ℓ_{conReg} by adopting the Jensen-Shannon (JS) divergence [33] $\mathbb{D}_{JS}$ to measure the similarity between the posterior probability distributions of the two classifiers as follows:

$$\begin{aligned}\ell_{conReg}(\theta_{C_1}, \theta_{C_2}) = {} & \mathbb{E}_{\mathrm{x}\sim p_u}\big[\mathbb{D}_{JS}\big(p_{C_1}(\mathrm{x}), p_{C_2}(\mathrm{x})\big)\big] \\ & + \mathbb{E}_{(z,\tilde{y})\sim p_g}\big[\mathbb{D}_{JS}\big(p_{C_1}(G(z,\tilde{y})), p_{C_2}(G(z,\tilde{y}))\big)\big].\end{aligned} \tag{8}$$

As a symmetrized and smoothed version of the KL divergence, $\mathbb{D}_{JS}$ is defined by

$$\begin{aligned}\mathbb{D}_{JS}\big(p_{C_1}(\mathrm{x}), p_{C_2}(\mathrm{x})\big) = {} & \frac{1}{2}\mathtt{KL}\big(p_{C_1}(\mathrm{x}) \| \bar{p}_C(\mathrm{x})\big) \\ & + \frac{1}{2}\mathtt{KL}\big(p_{C_2}(\mathrm{x}) \| \bar{p}_C(\mathrm{x})\big),\end{aligned} \tag{9}$$

In addition, $\mathbb{D}_{JS}(p_{C_1}(G(z,\tilde{y})), p_{C_2}(G(z,\tilde{y})))$ has a similar definition. Based on the definitions in Eq.(4) and Eqs.(8-9), minimizing ℓ_{conReg} leads to the classifiers producing predictions consistent with the consensus result $\bar{p}_C$.

The classifiers attempt to produce the predicted data pair $(\mathrm{x}, \bar{\mathrm{y}})$ for fooling the discriminator, where $\bar{\mathrm{y}}$ denotes the one-hot label determined by $\bar{p}_C(\mathrm{x})$. We need an adversarial training term for optimizing the classifiers, and the final formulation can be expressed as follows:

$$\begin{aligned}\min_{C_1,C_2} \; & \frac{1}{2}\mathbb{E}_{\mathrm{x}\sim p_u}\big[\bar{p}_C(\mathrm{x}) \log(1 - D(\mathrm{x}, \bar{\mathrm{y}}))\big] \\ & + \ell_{classify}(\theta_{C_1}) + \lambda\ell_{smoReg}(\theta_{C_1}) \\ & + \ell_{classify}(\theta_{C_2}) + \lambda\ell_{smoReg}(\theta_{C_2}) \\ & + \mu\ell_{conReg}(\theta_{C_1}, \theta_{C_2}),\end{aligned} \tag{10}$$

where λ and μ are the weighting factors for achieving a balance among the terms.

3.3. Adversarial Training

Since we follow the adversarial training scheme of the TripleGAN model in general, the discriminator D learns to distinguish the labeled data pair (x, y) from the synthesized data pair $(G(z), \tilde{y})$ and predicted data pair $(\mathrm{x}, \bar{\mathrm{y}})$. The corresponding optimization formulation is presented as follows:

$$\begin{aligned}\max_{D} \; & \mathbb{E}_{(x,y)\sim p_l}\big[\log D(x,y)\big] \\ & + \frac{1}{2}\mathbb{E}_{(z,\tilde{y})\sim p_g}\big[\log(1 - D(G(z,\tilde{y}),\tilde{y}))\big] \\ & + \frac{1}{2}\mathbb{E}_{\mathrm{x}\sim p_u}\big[\log(1 - D(\mathrm{x}, \bar{\mathrm{y}}))\big].\end{aligned} \tag{11}$$

The discriminator competes with the generator and classifiers in the minimax game. The generator attempts to syn-

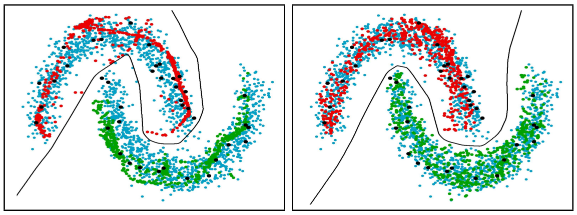

(a) TripleGAN* (b) EnhancedTGAN

Figure 3. Results on a toy examples for the baseline model and the proposed model. Different colors denote different types of data points: dark (labeled), cyan (unlabeled), and red/green (synthesized). The solid lines denote the resulting decision boundaries.

thesize high-fidelity instances, and the classifiers try to produce more accurate predictions on the unlabeled instances. When training the four modules jointly, collaborative learning between classifiers is able to provide more accurate categorical information of unlabeled data, which is crucial for the generator to learn the class-conditional distribution of the real data. More synthesized instances with high-fidelity can in turn be leveraged to improve classifier training, which lead to better decision boundaries and more accurate guidance to the generator. Therefore, the proposed EnhancedTGAN model is able to improve both instance synthesis and classification in the semi-supervised setting. The details of the corresponding optimization process are summarized in Algorithm 1.

4. Experiments

In this section, we verify the effectiveness of the proposed EnhancedTGAN model in semi-supervised instance synthesis and classification on both synthetic and real object recognition datasets. For a fair comparison with our baseline model TripleGAN [17], we implement this model in our configuration environment using the same setting as our EnhancedTGAN, and the resulting model is referred to as TripleGAN*. We also compare EnhancedTGAN with the state-of-the-art semi-supervised learning methods on multiple widely used benchmarks, including MNIST [16], SVHN [25] and CIFAR-10 [13]. Furthermore, we test the proposed model on FaceScrub [26] to investigate the quality of the synthesized human face images. In all the experiments, we perform labeled instance sampling 10 times, and report the mean and standard deviation of the test error rates for the classification task. In the class-conditional instance synthesis task, we present the synthesized images in a way that each group contains one image for each class and all of them share the same random vector.

4.1. Synthetic dataset

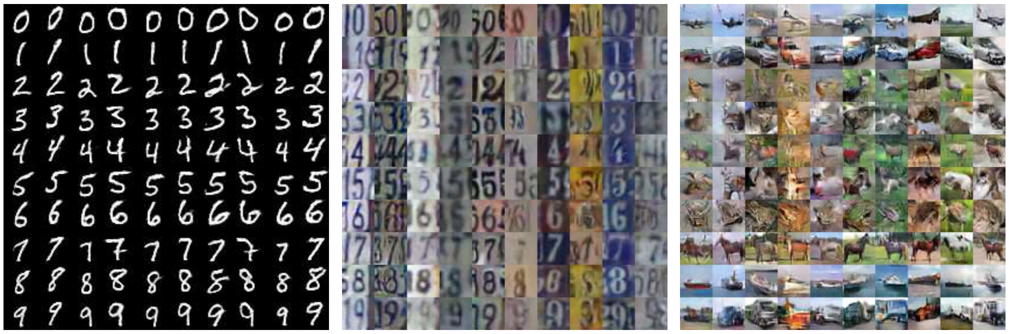
To show the effectiveness of our proposed feature-semantics matching approach, we compare TripleGAN* and EnhancedTGAN in terms of their capability to learn the class-conditional data distributions of a toy example. We adopt the 'two moons' synthetic dataset as shown in Figure 3, in which there are two classes, and each of them consists of 10 labeled data points and 1000 unlabeled data points. The generator, discriminator and classifiers are multi-layer perceptrons with 2-3 hidden layers. The two competing models share the same settings, but the proposed model has one more classifier than TripleGAN*. We train each model until it converges. The synthesized data points of TripleGAN* and EnhancedTGAN are shown in Figure 3(a) and (b), respectively. We use different colors (red and green) to denote the two classes of the synthesized data points. We can observe that the data points synthesized by TripleGAN* only lie in a portion of the real data distribution, and our EnhancedTGAN correctly learns the real data distributions. In addition, the decision boundary of the proposed model aligns better than that of the baseline.

4.2. Benchmark datasets

We further compare EnhancedTGAN with state-or-the-art semi-supervised deep learning models on the MNIST, SVHN and CIFAR-10 benchmarks, which are widely used for evaluation of classification and synthesis. According to the common setting, we perform experiments for the cases in which there are 100, 1000 and 4000 randomly selected labeled instances for MNIST, SVHN and CIFAR-10, respectively. The network architecture of the classifiers in EnhancedTGAN is the same as that in the main competing methods, such as TripleGAN and CT-GAN. The classification results are presented in Table 1. The error rates of the competing methods are taken from the existing literature, except TripleGAN*. TripleGAN* is a strong baseline, and outperforms the original TripleGAN. In all the cases, the proposed EnhancedTGAN achieves more accurate classification results than TripleGAN*. For CIFAR-10 with 4000 labels, EnhancedTGAN surpasses TripleGAN* by a large margin, and significantly reduces the test error rate from 14.65% to 9.42%. Compared with other competing methods, the proposed EnhancedTGAN produces more accurate or comparable classification results in all cases. Figure 4 shows the synthesized samples by our EnhancedTGAN model for the three datasets. We also visualize the t-SNE embedding [20] of the features associated with the last hidden layer of the classifier network in TripleGAN* and our EnhancedTGAN model on CIFAR-10 with 4000 labels. As shown in Figure 5, EnhancedTGAN performs better than TripleGAN* in learning the class-conditional data distributions, since we can observe that the samples are strongly clustered, and the distribution of the synthesized data can match the distribution of the unlabeled data very well.

Table 1. Comparison between our model and the competing methods on semi-supervised classification on the benchmark datasets.

	Test error rate (%) with # labels					
	MNIST		SVHN		CIFAR-10	
Method	100 labels	All labels	1000 labels	All labels	4000 labels	All labels
LadderNetwork[30]	1.06±0.37	0.57±0.02	-	-	20.40±0.47	-
SPCTN[40]	1.00±0.11	-	7.37±0.30	-	14.17±0.27	-
Π-model[15]	0.89±0.15	-	4.82±0.17	2.50±0.07	12.36±0.31	6.06±0.11
Temporal-Ensembling[15]	-	-	4.42±0.16	2.74±0.06	12.16±0.24	5.60±0.10
Mean-Teacher[36]	-	-	3.95±0.19	2.50±0.05	12.31±0.28	5.94±0.15
VAT[24]	-	-	3.74±0.09	2.69±0.04	11.96±0.10	5.65±0.17
VAdD[28]	-	-	4.16±0.08	2.31±0.01	11.68±0.19	5.27±0.10
VAdD+VAT[28]	-	-	3.55±0.05	**2.23±0.03**	10.07±0.11	**4.40±0.12**
SNTG+Π-model[19]	0.66±0.07	-	3.82±0.25	2.42±0.05	11.00±0.13	5.19±0.14
SNTG+VAT[19]	-	-	3.83±0.22	-	9.89±0.34	-
CatGAN[35]	1.39±0.28	-	-	-	19.58±0.58	-
Improved GAN[32]	0.93±0.07	-	8.11±1.30	-	18.63±2.32	-
ALI[5]	-	-	7.42±0.65	-	17.99±1.62	-
TripleGAN[17]	0.91±0.58	-	5.77±0.17	-	16.99±0.36	-
GoodBadGAN[4]	0.80±0.10	-	4.25±0.03	-	14.41±0.03	-
CT-GAN[38]	0.89±0.13	-	-	-	9.98±0.21	-
TripleGAN*	0.81±0.08	0.31±0.04	4.53±0.22	2.94±0.15	14.65±0.38	6.64±0.13
EnhancedTGAN	**0.42±0.03**	**0.27±0.03**	**2.97±0.09**	**2.23±0.01**	**9.42±0.22**	4.80±0.07



(a) MNIST (b) SVHN (c) CIFAR-10

Figure 4. Synthesized instances produced by the proposed EnhancedTGAN model for semi-supervised class-conditional object image synthesis on MNIST with 100 labels, SVHN with 1000 labels and CIFAR-10 with 4000 labels. Each row has the same class label, and each column is synthesized from the same random vector.

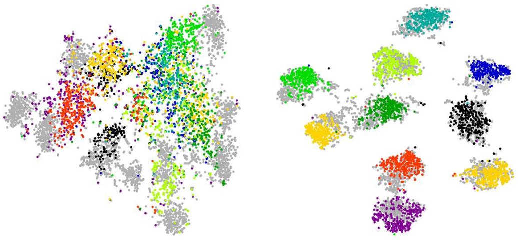
Table 2. Ablation study of the proposed model on CIFAR-10 with 4000 labels for investigating the influence of synthesized data, and consistency and smoothness regularization in semi-supervised classification.

	Test error rate (%) with # labels
Method	4000 labels
w/o GAN	11.92±0.19
w/o ℓ_{conReg}	11.47±0.13
w/o ℓ_{smoReg}	12.03±0.29
EnhancedTGAN	**9.42±0.22**

4.3. Ablation study

We remove the feature-semantics matching terms from the overall loss function of the generator on CIFAR-10 with 4000 labels, and show the classification accuracy on the synthesized data during training in Figure 6. We can observe that the class-wise mean feature matching term is able to boost the classification accuracy, which indicates that the synthesized data can better match the statistics of the real data. The semantics matching term can further improve the training of the generator.

To investigate the effectiveness of our proposed improvement strategies in semi-supervised classification, we perform an ablation study to compare the resulting models when removing the corresponding modules on CIFAR-10 with 4000 labels as shown in Table 2. We first remove the generator and the discriminator to evaluate the classifiers with collaborative learning, and the test error rate rises to 11.92%, which indicates that the synthesized instances are useful for improving classifier training. In addition, we remove the consistency and smoothness regularization terms from the overall loss function of the classifiers to investigate the influence of collaborative learning and the local Lipschitz condition, and a significant performance drop can be observed in both cases. We consider that both synthesized data and regularization are important for improving semi-



(a) TripleGAN* (b) EnhancedTGAN

Figure 5. The t-SNE embedding of unlabeled training data and synthesized data on CIFAR-10 with 4000 labels. The unlabeled samples are marked gray, and different classes of synthesized samples are marked different colors.

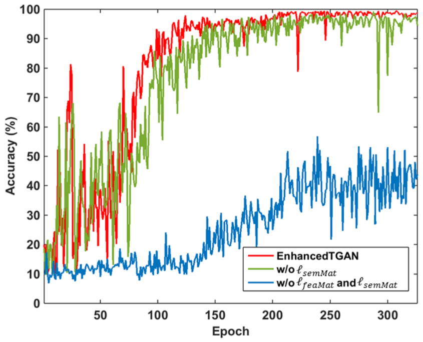
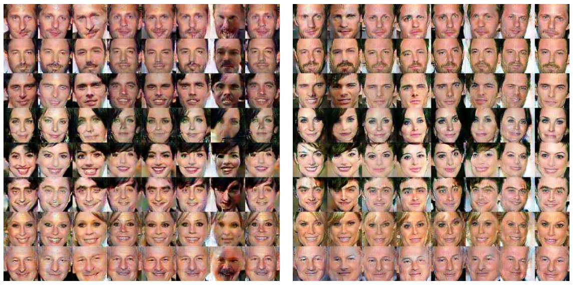


Figure 6. Classification accuracy on synthesized data when removing the feature-semantics matching terms from the overall loss function of the generator on CIFAR-10 with 4000 labels.

supervised classification.

4.4. Face synthesis

To further investigate the capability of the proposed EnhancedTGAN model in performing difficult semi-supervised instance synthesis, we conduct an experiment on the FaceScrub dataset. Since the classes in this dataset contain different numbers of human face images, we select the 100 largest classes in our experiment, and only 20 images sampled randomly in each class are labeled. All the images are resized to 64 × 64, and thus we slightly modify the network architectures used previously for this experiment without significantly increasing the number of model parameters. The synthesized human face images are shown in Figure 7. For TripleGAN*, we observe that the variation within a class is relatively small, and the structures of the human faces are lost in some images. On the other hand, the synthesized images of our EnhancedTGAN look realistic and preserve human identities. The corresponding classification results of TripleGAN* and EnhancedTGAN are shown in Table 3.



(a) TripleGAN* (b) EnhancedTGAN

Figure 7. Synthesized instances produced by the TripleGAN* and EnhancedTGAN models for semi-supervised class-conditional human face image synthesis on FaceScrub with 2000 labels. Each row has the same class label, and each column is synthesized from the same random vector.

Table 3. Comparison between the baseline model and the proposed model on semi-supervised classification on FaceScrub-100.

	Test error rate (%) with # labels	
Method	2000 labels	All labels
TripleGAN*	18.23±0.56	5.43±0.41
EnhancedTGAN	**16.08±0.24**	**4.29±0.20**

5. Conclusion

In this paper, we propose an enhanced TripleGAN model for improving both semi-supervised conditional instance synthesis and classification. Toward this end, we adopt feature-semantics matching to force the generator to effectively learn the class-conditional data distributions, such that the synthesized instances with high-fidelity can be used for training better classifiers. On the other hand, we collaboratively train two classifiers, which can provide more accurate guidance for the generator. The experiment results demonstrate that the proposed model outperforms the original TripleGAN and achieves new state-of-the-art results on multiple benchmark datasets.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Project No. 61502173, U1611461, 61722205, 61751205, 61572199), in part by the Research Grants Council of the Hong Kong Special Administration Region (Project No. CityU 11300715), in part by City University of Hong Kong (Project No. 7005055), in part by the Natural Science Foundation of Guangdong Province (Project No. 2016A030310422), in part by Key R&D Program of Guangdong Province (Project No. 2018B010107002), and in part by the Fundamental Research Funds for the Central Universities (Project No. 2018ZD33).

References

[1] M. Abbasnejad, A. Dick, and A. Hengel. Infinite variational autoencoder for semi-supervised learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, pages 781 – 790, 2017.

[2] M. Arjovsky, S. Chintala, and L. Bottou. Wasserstein generative adversarial networks. In *Proc. International Conference on Machine Learning*, pages 214 – 223, 2017.

[3] J. Bao, D. Chen, F. Wen, H. Li, and G. Hua. CVAE-GAN: fine-grained image generation through asymmetric training. In *Proc. International Conference on Computer Vision*, pages 2745 – 2754, 2017.

[4] Z. Dai, Z. Yang, F. Yang, W. Cohen, and R. Salakhutdinov. Good semi-supervised learning that requires a bad GAN. In *Proc. Advances in Neural Information Processing Systems*, pages 6513 – 6523, 2017.

[5] V. Dumoulin, I. Belghazi, B. Poole, O. Mastropietro, A. Lamb, M. Arjovsky, and A. Courville. Adversarially learned inference. In *Proc. International Conference on Learning Representation*, 2017.

[6] G. Dziugaite, D. Roy, and Z. Ghahramani. Training generative neural networks via maximum mean discrepancy optimization. In *Proc. Conference on Uncertainty in Artificial Intelligence*, pages 258–267, 2015.

[7] Z. Gan, L. Chen, W. Wang, Y. Pu, Y. Zhang, H. Liu, C. Li, and L. Carin. Triangle generative adversarial networks. In *Proc. Advances in Neural Information Processing Systems*, 2017.

[8] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Proc. Advances in Neural Information Processing Systems*, pages 2672 – 2680, 2014.

[9] D. Kingma and J. Ba. Adam: a method for stochastic optimization. In *Proc. International Conference on Learning Representations*, 2015.

[10] D. Kingma, S. Mohamed, D. Rezende, and M. Welling. Semi-supervised learning with deep generative models. In *Proc. Neural Information Processing Systmes*, pages 3581 – 3589, 2017.

[11] D. Kingma and M. Welling. Auto-encoding variational Bayes. In *Proc. International Conference on Learning Representation*, 2014.

[12] T. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *Proc. International Conference on Learning Representation*, 2017.

[13] A. Krizhevsky and G. Hinton. Learning multiple layers of features from tiny images. In *Technical Report*, 2009.

[14] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Proc. Neural Information Processing Systmes*, pages 1106 – 1114, 2014.

[15] S. Laine and T. Aila. Temporal ensembling for semi-supervised learning. In *Proc. International Conference on Learning Representations*, 2017.

[16] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278 – 2324, 1998.

[17] C. Li, K. Xu, J. Zhu, and B. Zhang. Triple generative adversarial nets. In *Proc. Advances in Neural Information Processing Systems*, pages 1195 – 1204, 2017.

[18] C. Li, J. Zhu, and B. Zhang. Max-margin deep generative models for (semi-)supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(11):2762 – 2775, 2018.

[19] Y. Luo, J. Zhu, M. Li, Y. Ren, and B. Zhang. Smooth neighbors on teacher graphs for semi-supervised learning. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[20] L. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(11):2579 – 2605, 2008.

[21] T. Miyato, T. Kataoka, M. Koyama, and Y. Yoshida. Spectral normalization for generative adversarial networks. In *Proc. International Conference on Learning Representation*, 2018.

[22] T. Miyato and M. Koyama. cGANs with projection discriminator. In *Proc. International Conference on Learning Representation*, 2018.

[23] T. Miyato, S. Maeda, S. Ishii, and M. Koyama. Virtual adversarial training: a regularization method for supervised and semi-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence (Early Access)*, 2018.

[24] T. Miyato, S. Maeda, M. Koyama, K. Nakae, and S. Ishii. Distributional smoothing with virtual adversarial training. In *Proc. International Conference on Learning Representations*, 2016.

[25] Y. Netzer, T. Wang, A. Goates, A. Bissacco, B. Wu, and A. Ng. Reading digits in natural images with unsupervised feature learning. In *Proc. NIPS Workshop on Deep Learning and Unsupervised Feature Learning*, 2011.

[26] H. Ng and S. Winkler. A data-driven approach to cleaning large face datasets. In *Proc. International Conference on Image Processing*, pages 343 – 347, 2014.

[27] A. Odena. Semi-supervised learning with generative adversarial networks. In *Proc. International Conference on Learning Representation*, 2016.

[28] S. Park, J. Park, S. Shin, and I. Moon. Adversarial dropout for supervised and semi-supervised learning. In *Proc. AAAI Conference on Artificial Intelligence*, 2018.

[29] A. Radford, L. Metz, and S. Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. In *Proc. International Conference on Learning Representation*, 2016.

[30] A. Rasmus, M. Berglund, M. Honkala, H. Valpola, and T. Raiko. Semi-supervised learning with ladder networks. In *Proc. Neural Information Processing Systmes*, pages 3546 – 3554, 2015.

[31] S. Ren, K. He, R. Girshick, and J. Sun. Faster R-CNN: Towards real-time object detection with region proposal networks. In *Proc. Advances in Neural Information Processing Systems*, 2015.

[32] T. Salimans, I. Goodfellow, W. Zaremba, and V. Cheung. Improved techniques for training GANs. In *Proc. Neural Information Processing Systmes*, pages 2234 – 2242, 2016.

[33] H. Schutze and C. Manning. *Fundations of statistical nutural language processing*. MIT Press, Cambridge, Mass, 1999.

[34] E. Shelhamer, J. Long, and T. Darrell. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):640 – 651, 2017.

[35] J. Springenberg. Unsupervised and semi-supervised learning with categorical generative adversarial networks. In *Proc. International Conference on Learning Representations*, 2016.

[36] A. Tarvainen and H. Valpola. Mean teachers are better role models: weight-averaged consistency targets improve semi-supervised deep learning results. In *Proc. Advances in Neural Information Processing Systems*, 2017.

[37] H. Valpola. From neural PCA to deep unsupervised learning. In *Proc. Advances in Independent Component Analysis and Learning Machines, arXiv:1411.7783*, 2015.

[38] X. Wei, B. Gong, Z. Liu, W. Lu, and L. Wang. Improving the improved training of Wasserstein GANs: a consistency term and its dual effect. In *Proc. International Conference on Learning Representations*, 2018.

[39] H. Wu and S. Prasad. Semi-supervised deep learning using pseudo labels for hyperspectral image classification. *IEEE Transactions on Image Processing*, 27(3):1259 – 1270, 2018.

[40] S. Wu, Q. Ji, S. Wang, H. Wong, Z. Yu, and Y. Xu. Semi-supervised image classification with self-paced cross-task networks. *IEEE Transactions on Multimedia*, 20(4):851–865, 2018.

[41] Z. Yang, W. Cohen, and R. Salakhutdinov. Revisiting semi-supervised learning with graph embeddings. In *Proc. International Conference on Machine Learning*, 2016.

[42] T. Young, D. Hazarika, S. Poria, and E. Cambria. Recent trends in deep learning based natural language processing. *IEEE Computational Intelligence Magazine*, 13(3):55 – 75, 2018.

[43] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. In *arXiv preprint arXiv:1705.05512*, 2018.