EECE7360 final project

GAN 图像风格转化 – style transfer

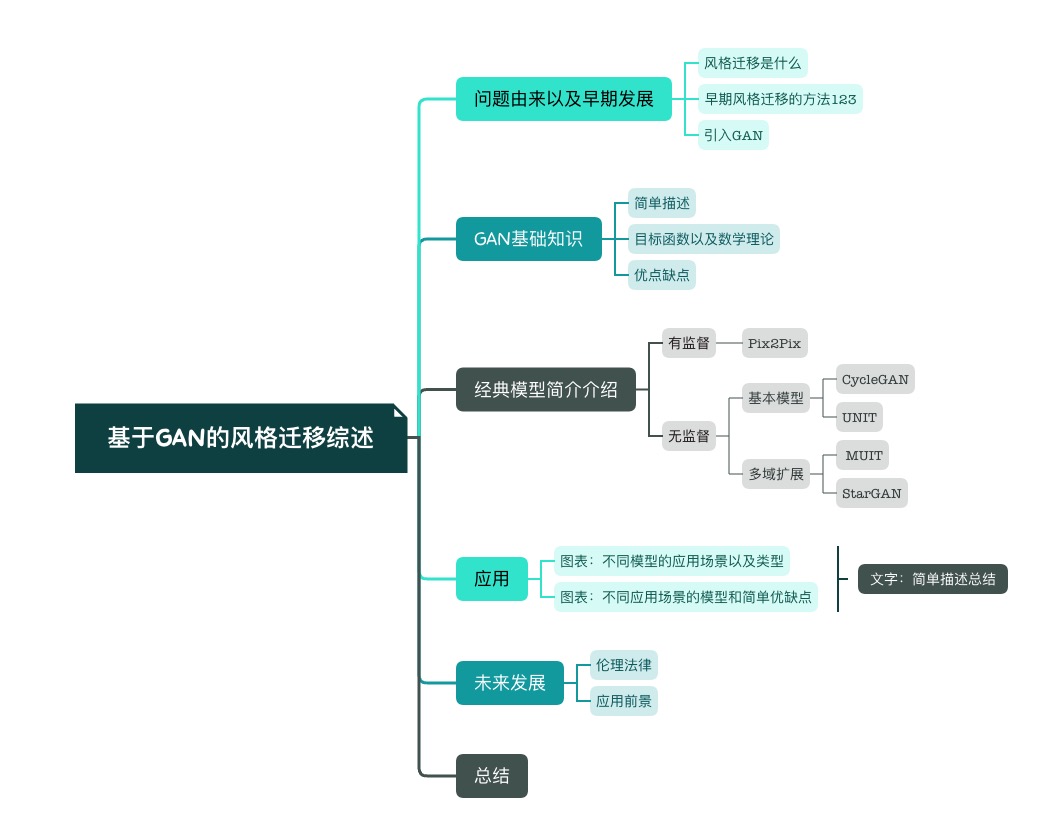
# 重要paper

Generative Adversarial Networks（arxiv：<https://arxiv.org/abs/1406.2661）GAN>开山之作

CycleGAN

基于 GAN 网络的风格迁移方法(生成领域的王者，生成质量很好，但训练起来很难，超级多的工程trick)

发展历程：Goodfellow大神首次提出GAN的生成模型（开山之作）->CGAN->CycleGAN(经典之作，解决了pair数据难收集的问题)->UNIT(结合了GAN和另一个生成大哥VAE,适用于真实图像间的相互转换)->StarGAN(多领域转换)->CartoonGAN(开始整卡通了,加入了边缘对抗损失和内容损失,关注动漫场景)->SCGAN(关注点放在拍摄者面部)->ACL-GAN（与CycleGAN的区别在于，ACL-GAN不要求生成图像能够完全翻译回源图像，只需要保留源图像的重要信息即可，也即只需要在特征层面做约束）->CariGAN（关注卡通图像的几何形变）



Graphical user interface

Description automatically generated with low confidence

Unsupervised Image-to-image Translation Networks (UNIT)

# Introduction

Unsupervised image-to-image translation aims at learning a joint distribution of images in different domains by using images from the marginal distributions in individual domains.

make a shared-latent space assumption and propose an unsupervised image-to-image translation framework based on Coupled GANs

compare the proposed framework with competing approaches and present high quality image translation results on various challenging unsupervised image translation tasks

UNIT framework that are based on generative adversarial networks (GANs) and variational autoencoders (VAEs)

## unsupervised

In the unsupervised setting, we only have two independent sets of images where one consists of images in one domain and the other consists of images in another domain—there exist no paired examples showing how an image could be translated to a corresponding image in another domain.

## Key challenge

learn a joint distribution of images in different domains

The coupling theory [16] states there exist an infinite set of joint distributions that can arrive the given marginal distributions in general. Hence, inferring the joint distribution from the marginal distributions is a highly ill-posed problem.

need additional assumptions on the structure of the joint distribution.

– shared-latent space assumption

assumes a pair of corresponding images in different domains can be mapped to a same latent representation in a shared-latent space.

# Assumptions

In unsupervised image-to-image translation, we are given samples drawn from the marginal distributions PX1 (x1) and PX2 (x2).

we assume for any given pair of images x1 and x2, there exists a shared latent code z in a shared-latent space, such that we can recover both images from this code, and we can compute this code from each of the two images.

shared-latent space assumption implies the cycle-consistency assumption

To implement the shared-latent space assumption, we further assume a shared intermediate representation h such that the process of generating a pair of corresponding images admits a form of

Black text on a white background

Description automatically generated

# Framework

Diagram

Description automatically generated

based on variational autoencoders (VAEs) [13, 22, 14] and generative adversarial networks (GANs) [6, 17]

It consists of 6 subnetworks: including two domain image encoders E1 and E2, two domain image generators G1 and G2, and two domain adversarial discriminators D1 and D2.

## VAE

The encoder–generator pair {E1, G1} constitutes a VAE for the X1 domain, termed VAE1. For

an input image x1 2 X1, the VAE1 first maps x1 to a code in a latent space Z via the encoder E1 and then decodes a random-perturbed version of the code to reconstruct the input image via the generator G1.

## Weight-sharing

Based on the shared-latent space assumption discussed in Section 2, we enforce a weight-sharing constraint to relate the two VAEs. Specifically, we share the weights of the last few layers of E1 and E2 that are responsible for extracting high-level representations of the input images in the two domains. Similarly, we share the weights of the first few layers of G1 and G2 responsible for decoding high-level representations for reconstructing the input images.

In the unsupervised setting, no pair of corresponding images in the two domains exists to train the network to output a same latent code.

## GAN

Our framework has two generative adversarial networks: GAN1 = {D1 , G1 } and GAN2 =

{D2, G2}. In GAN1, for real images sampled from the first domain, D1 should output true, while

for images generated by G1, it should output false.

G1 can generate two types of images: 1) images from the reconstruction stream x ̃1!1 = G (z ⇠ q (z |x )) and 2) images from the translation 111111 stream x ̃2!1 = G (z ⇠ q (z |x )). Since the reconstruction stream can be supervisedly trained, it 212222 is suffice that we only apply adversarial training to images from the translation stream, x ̃2!1. We 2 apply a similar processing to GAN2

# Experiment

Chart, line chart

Description automatically generated

We first analyze various components of the proposed framework. We then present visual results on challenging translation tasks. Finally, we apply our framework to the domain adaptation tasks.

In one experiment, we varied the number of weight-sharing layers in the VAEs and paired each configuration with discriminator architectures of different depths during training. We changed the number of weight-sharing layers from 1 to 4.

The results were reported in Figure 2(b). Each curve corresponded to a discriminator architecture of a different depth. The x-axis denoted the number of weigh-sharing layers in the VAEs. We found that the shallowest discriminator architecture led to the worst performance. We also found that the number of weight-sharing layer had little impact. This was due to the use of the residual blocks.

We analyzed impact of the hyper-parameter values to the translation accuracy. For different weight values on the negative log likelihood terms (i.e., 2, 4), we computed the achieved translation accuracy over different weight values on the KL terms (i.e., 1, 3). The results were reported in Figure 2(c). We found that, in general, a larger weight value on the negative log likelihood terms yielded a better translation accuracy. We also found setting the weights of the KL terms to 0.1 resulted in consistently good performance. We hence set 1 = 3 = 0.1 and 2 = 4 = 100.

# Conclusion

The current framework has two limitations. First, the translation model is unimodal due to the Gaussian latent space assumption. Second, training could be unstable due to the saddle point searching problem. We plan to address these issues in the future work.

MUNIT

# Introduction

assume that the image representation can be decomposed into a content code that is domain-invariant, and a style code that captures domain-specific properties. To translate an image to another domain, we recombine its content code with a random style code sampled from the style space of the target domain. We analyze the proposed framework and establish several theoretical results.

In many scenarios, the cross-domain mapping of interest is multimodal.

Diagram

Description automatically generated

# Assumption

latent space of images can be decomposed into a content space and a style space.

images in different domains share a common content space but not the style space

# Framework

Chart

Description automatically generated

StarGAN

# Introduction

Recent studies have shown remarkable success in image- to-image translation for two domains. However, existing approaches have limited scalability and robustness in han- dling more than two domains, since different models should be built independently for every pair of image domains. To address this limitation, we propose StarGAN, a novel and scalable approach that can perform image-to-image translations for multiple domains using only a single model.

However, existing models are both inefficient and ineffective in such multi-domain image translation tasks. Their inefficiency results from the fact that in order to learn all mappings among k domains, k(k−1) generators have to be trained.

## Dataset

Several image datasets come with a number of labeled attributes. For instance, the CelebA[19] dataset contains 40 labels related to facial attributes such as hair color, gender, and age, and the RaFD [13] dataset has 8 labels for facial expressions such as ‘happy’, ‘angry’ and ‘sad’. These set- tings enable us to perform more interesting tasks, namely multi-domain image-to-image translation, where we change images according to attributes from multiple domains.

# Contribution

* We propose StarGAN, a novel generative adversarial network that learns the mappings among multiple do- mains using only a single generator and a discrimina- tor, training effectively from images of all domains.
* We demonstrate how we can successfully learn multi- domain image translation between multiple datasets by utilizing a mask vector method that enables StarGAN to control all available domain labels.
* We provide both qualitative and quantitative results on facial attribute transfer and facial expression synthesis tasks using StarGAN, showing its superiority over baseline models.

# Framework

Diagram

Description automatically generated

Fig. 2 (a) illustrates how twelve distinct generator networks have to be trained to translate images among four different domains. Meanwhile, they are ineffective that even though there exist global features that can be learned from images of all domains such as face shapes, each generator cannot fully utilize the entire training data and only can learn from two domains out of k. Failure to fully utilize training data is likely to limit the quality of generated images. Furthermore, they are incapable of jointly training domains from different datasets because each dataset is partially labeled.

As a solution to such problems we propose StarGAN, a novel and scalable approach capable of learning mappings among multiple domains.

model takes in training data of multiple domains, and learns the mappings between all available domains using only a single generator.

our generator takes in as inputs both image and domain information, and learns to flexibly translate the image into the correspond- ing domain. We use a label (e.g., binary or one-hot vector) to represent domain information. During training, we ran- domly generate a target domain label and train the model to flexibly translate an input image into the target domain. By doing so, we can control the domain label and translate the image into any desired domain at testing phase.

enables joint training between domains of different datasets by adding a mask vector to the domain label. Our proposed method ensures that the model can ignore unknown labels and focus on the label provided by a particular dataset

Graphical user interface

Description automatically generated with medium confidence

train a single generator G that learns mappings among multiple domains

# Advantages

An important advantage of StarGAN is that it simulta- neously incorporates multiple datasets containing different types of labels, so that StarGAN can control all the labels at the test phase.

Mask Vector. To alleviate this problem, we introduce a mask vector m that allows StarGAN to ignore unspecified labels and focus on the explicitly known label provided by a particular dataset.

# Disadvantages

An issue when learning from multiple datasets, however, is that the label information is only partially known to each dataset. In the case of CelebA [19] and RaFD [13], while the former contains labels for attributes such as hair color and gender, it does not have any labels for facial expressions such as ‘happy’ and ‘angry’, and vice versa for the latter. This is problematic because the complete information on the label vector c′ is required when reconstructing the input image x from the translated image G(x, c) (See Eq. (4)).

Mask Vector. To alleviate this problem, we introduce a mask vector m that allows StarGAN to ignore unspecified labels and focus on the explicitly known label provided by a particular dataset.

# Experiments

compare StarGAN against recent methods on facial attribute transfer by conducting user studies.

perform a classification experiment on facial expression synthesis.

发展总结

Liu, M. Y., Breuel, T., & Kautz, J. (2017). Unsupervised image-to-image translation networks. Advances in neural information processing systems, 30.

Huang, X., Liu, M. Y., Belongie, S., & Kautz, J. (2018). Multimodal unsupervised image-to-image translation. In Proceedings of the European conference on computer vision (ECCV) (pp. 172-189).

Choi, Y., Choi, M., Kim, M., Ha, J. W., Kim, S., & Choo, J. (2018). Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 8789-8797).