Auto-GAN: Self-Supervised Collaborative Learning for Medical Image Synthesis

Bing Cao,[1] Han Zhang,[2*] Nannan Wang,[1*] Xinbo Gao,[1] Dinggang Shen[2*]
[1]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an, China
[2]Department of Radiology and BRIC, University of North Carolina at Chapel Hill, Chapel Hill, USA

Abstract

In various clinical scenarios, medical image is crucial in disease diagnosis and treatment. Different modalities of medical images provide complementary information and jointly helps doctors to make accurate clinical decision. However, due to clinical and practical restrictions, certain imaging modalities may be unavailable nor complete. To impute missing data with adequate clinical accuracy, here we propose a framework called self-supervised collaborative learning to synthesize missing modality for medical images. The proposed method comprehensively utilize all available information correlated to the target modality from multi-source-modality images to generate any missing modality in a single model. Different from the existing methods, we introduce an autoencoder network as a novel, self-supervised constraint, which provides target-modality-specific information to guide generator training. In addition, we design a modality mask vector as the target modality label. With experiments on multiple medical image databases, we demonstrate a great generalization ability as well as specialty of our method compared with other state-of-the-arts.

Introduction

Multimodal medical imaging such as computed tomography (CT), magnetic resonance imaging (MRI), positron emission tomography (PET) plays an irreplaceable role in routine clinical practice (*e.g,* lesion detection, disease diagnosis, and treatment planning). They provide unprecedented *in vivo* resolution and contrast in visualization of internal organs of almost entire human body and the modern medicine, especially precision medicine, has increasingly depended on them.

In many clinical scenarios, doctors combine different insights from multi-modality medical images (*e.g.,* T1-weighted (T1), T2-weighted (T2) and T2-FLAIR (FLuid-Attenuated Inversion Recovery), and T1 post-contrast images) (T1-C) to make a final diagnostic decision. Each imaging modality provides certain information. While the doctors desire as many different modalities as possible to make a joint and more accurate decision, certain limitations (*e.g.,* restricted medical conditions, inadequate scanning time, and cost/spending/resource control) could result in sacrificing some imaging modalities to have the most important one done. For examples, low-resolution (*e.g.,* thick-sliced) imaging modalities are always combined with another single high-resolution imaging modality to save scanning time and the low-resolution images can be misleading due to its partial coverage. CT is notorious for poor soft-tissue contrast and radiation exposure, but can better quantify tissue's physical density. MRI is non-invasive and safer with better contrast; therefore, it could be more prioritized than CT with limited scanning time.

Due to the aforementioned restrictions on acquiring full multi-modality images, imputing the missing or low-quality modalities by using available modalities has become a very important research topic in the artificial intelligence-based medical image analysis (Van Buuren, Boshuizen, and Knook 1999; Sauerbrei and Royston 1999; Shen, Wu, and Suk 2017). Mounting effort has been put in developing effective image imputation algorithms for not only medical image analysis but also, more generally, natural image synthesis (*e.g.,* style transfer (Gatys, Ecker, and Bethge 2016), denoising (Im et al. 2017), and super-resolution (Ledig et al. 2017)), where a mapping function is learned to translate images from a source domain to a target domain. While classic machine learning techniques (Jog et al. 2017) have made enormous achievement, recently, generative adversarial networks (GANs) (Goodfellow et al. 2014) has shown unprecedented superiority in such image generation tasks, including its recent variation conditional GAN (conduct image-to-image translation by using an imposed condition), *e.g.,* Pix2Pix (Isola et al. 2017), CycleGAN (Zhu et al. 2017), and StarGAN (Choi et al. 2018). However, these methods only can transfer images from one to another domain, instead of translating multi-modality images from multiple domains to a target (new modality) domain. Since multi-modality images constitute fundamentally complementary information to each other (Yu et al. 2018; 2019), it is theoretically and practically necessary to use all the available information to generate more accurate missing modalities. In addition, the

[*]Corresponding authors: Nannan Wang (nnwang@xidian.edu.cn), Han Zhang (hanzhang@med.unc.edu), and Dinggang Shen (dgshen@med.unc.edu)

weak supervision and unguided generation/translation due to ignored feature-level constraint could lead to greater deformation and indistinct details in output images. For instance, the results could be heavily biased towards source modalities in the CycleGAN and greatly distorted in the Pix2Pix.

To address the drawbacks, we introduce an auto-encoder network with an encoder-decoder network architecture and take advantage of its strong self-representation ability (Hinton and Salakhutdinov 2006) to impose self-supervised feature-level constraint and better guide a target domain-specific generator. We coin our method as Auto-GAN, which jointly utilizes multiple source domains (multi-modality images) and self representation of the target domain to deeply and collaboratively supervise a better decoder in a layer-by-layer fashion. Our Auto-GAN consists of multiple branches as the encoder network anchored with multiple input modalities and another branch driven by the self-representation network as the decoder network. In-between these two networks, we use latent layers to fuse and distill multi-modality features and feed the decoder network. While our method is a general solution to various medical imaging modalities, it can be further empowered by a modality mask vector, which is utilized as the target modality label. By concatenating the label with input images, our model can translate any combinations of available modalities to the missing modality in a single model. There are many advantages of the Auto-GAN in comparing with the existing methods:

- The deep representations extracted from the auto-encoder can provide a strong self-representation ability to supervise image translation framework. Therefore, the proposed Auto-GAN can generate more accurate results.
- Multiple branches in the encoder network can incorporate collaborative information from multiple source domains to provide complementary structural details for the decoder network.
- By imposing a modality mask vector to the inputs, Auto-GAN can estimate different modalities with a single model, which is more effective than other state-of-the-art methods and achieves superior performance in both qualitative and quantitative evaluations.

Related Work

Existing medical image synthesis methods can be grouped into two categories: data-driven methods and model-driven methods.

Data-driven methods utilize training data in the testing phase, which results in the estimations deeply depends on the integrity and the accuracy of training data. The images in both training set and testing set are cropped into patches. For example, Burgos et al. (2014) proposed an information propagation scheme, which searches the most similar patches from the training data to reconstruct the corresponding patch from source domain to target domain. This method is further improved by Vemulapalli et al. (2015), which estimates each target voxel corresponding to the source domain by searching the nearest neighbor voxels from the training set in the target domain. These methods require paired cross-domain training data and a large scale of training set. However, large dataset leads to increased computational complexity and patch-based synthesis inevitably causes a blurring effect, reducing the quality of the generated images.

Model-driven methods utilize training data in generative model learning for mapping the image from source to target domain. Goodfellow et al. (2014) first proposed GANs to generate a target image without a certain input image, which is then improved to conditional generative adversarial networks (cGAN). Different from GANs, cGAN takes the input images into consideration. Isola et al. (2017) proposed Pix2Pix, a generative model that learns a mapping function from the paired cross-domain data and utilize a generative loss and a discriminator loss to constrain the networks. To apply the Pix2Pix model to unpaired data, Zhu et al. (2017) proposed CycleGAN, where a cycle-consistent loss was designed to reconstruct the input images from the generated images. These methods achieved better perceptual appearance and much improved details with less computational demands than the data-driven methods; however, there could be excessive deformation in the synthesized images, and this may affect their clinical applications.

Auto-GAN

To solve these issues, we propose an Auto-GAN framework and detail it as below. Auto-GAN can generate any missing modality from available modalities in a unified single model.

Motivation

Although existing GAN-based methods have largely improved the quality of synthesized images, these images are often found to be deformed and/or blurred. The main reasons are that these methods implement loss functions computed by the pixel-level difference between generated images and the ground-truth (cycle-consistent loss (Zhu et al. 2017)) or the discriminator loss (as used in "PatchGAN" (Isola et al. 2017)). To our best knowledge, no work uses a feature-level constraint directly to guide the decoder for a better learned generator.

Inspired by knowledge distillation (Hinton, Vinyals, and Dean 2015; Kim and Rush 2016; Liu et al. 2019), which extracts general, moderate and sufficient knowledge from a "teacher" network to guide the "student" network, we introduce an experienced teacher network to guide the decoder in the generative network at the feature level. To better guide the decoder, a network with a strong representation ability is required. To this end, the classification models (Krizhevsky, Sutskever, and Hinton 2012; Simonyan and Zisserman 2014; He et al. 2016) can be pre-trained on large-scale natural image datasets (Deng et al. 2009) to extract sufficient feature maps with a strong representation ability as knowledge transfer. However, for medical images that are more complex compared to natural images (Huang et al. 2019), it is difficult to directly borrow the natural image-derived knowledge for the guidence of generator networks. In practice, it is impossible to acquire large-scale medical image data sets for pre-training as well. Taken together, it should be better

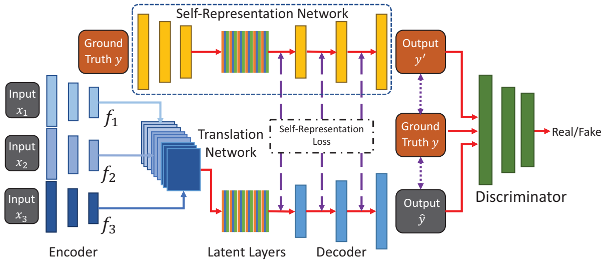

Figure 1: The framework of Auto-GAN. There are three major components: Self-Representation Network, Translation Network, and Discriminator Network.

dealt with for medical image synthesis than natural image synthesis.

Auto-encoder network can be supervised by the input images themselves and has a similar architecture to that of the generator in the GAN-based methods. Since the auto-encoder only works on a single domain, it is also easy to converge (faster than the generators that learn the mapping function between two different domains). Its strong self-representation ability has made it widely used in other tasks (*e.g.*, feature dimensionality reduction). We hereby borrow auto-encoder network to guide the decoder network at the feature level, which is better than solely learning from the reconstructed images through back-propagation with a pixel-level loss.

Self-Supervised Collaborative Learning

The framework of the proposed Auto-GAN consists of three key components: a translation network T (to translate images from source to target domain) , a self-representation network S (to guide the decoder), and a discriminator network D, as shown in Fig. 1. The three components are trained in an end-to-end fashion. For translation network T, there are multiple branches (each for one modality) in the encoder network, six latent layers, and one branch in the decoder network. Different from the translation network T, there are only one branch in both encoder and decoder of the self-representation network S (for the output modality).

In the training phase, the translation network T encodes the input images into a common latent feature space. The latent layers fuses the concatenated deep features from the input images to extract their complementary information for the image generation through a decoder. The self-representation network takes a form of an auto-encoder and is trained by the target images only. Once well trained, we use the feature maps extracted from the decoder of the self-representation network S to guide the optimization of the decoder of the translation network T.

In the testing phase, the self-representation network and discriminator network are removed, only the translation network T is used to translate the images from multiple source domains to the target domain. For different input combinations from different source domains, our framework can generate the images of the missing modalities by a single, unified model.

Implementation

Without loss of generality and for ease of representation, we assume four modalities $\{m_1, m_2, m_3, m_4\}$ in the data sets.

Translation Network We exploit an encoder-decoder network architecture for the translation network T, inspired by existing image translation methods (Noh, Hong, and Han 2015; Justin, Alexandre, and Li 2016; Long, Shelhamer, and Darrell 2017). The translation network T consists of a multi-branch encoder EC, a latent network L, and a decoder DC. The number of branches in the encoder network is determined by the total number of input modalities. Each branch has three convolutional layers, followed by a batch-normalization (BN) layer (Ioffe and Szegedy 2015). The latent network consists of six residual blocks (He et al. 2016), each of which takes a form of Conv-BN-ReLu-Conv-BN. For each residual block, the input is skip-connected with the output of the last batch-normalization layer.

For the given input images $\{x_{m_1}, x_{m_2}, x_{m_3}\}$ of modalities $\{m_1, m_2, m_3\}$, the branches $EC_i^T, (i \in \{1, 2, 3\})$ of the encoder in the translation network T encode the input images into a common latent feature space as $\{f_1^G, f_2^G, f_3^G\}$,

$$f_i^G = EC_i^T(x_{m_i}), i \in \{1, 2, 3\} \tag{1}$$

where $EC^T(\cdot)$ denotes the forward computation process of the convolution network, and i denotes modality. Then, the latent layers L extract the fused complementary information f^G from the concatenated encoded features as $f^G = L(f_1^G, f_2^G, f_3^G)$. The decoder DC^T extracts the feature maps $f^{G,DC}$ from f^G as:

$$f_i^{G,DC} = DC^{T,i}(f^G) \tag{2}$$

where i denotes the i-th layer of decoder network. $DC^{T,i}(\cdot)$ denotes the forward computation process of the decoder in T.

Self-representation network As aforementioned, an auto-encoder is trained to reconstruct the input itself, which ensures a strong representation ability for the auto-encoder in the same domain. Therefore, Auto-GAN takes the auto-encoder network as a self-representation network. Considering this, one of the key concepts for the proposed framework is guiding the decoder in translation network by the decoder in a self-representation network. Here, we utilize the same network architecture as the translation network T for the self-representation network S, except merging multiple branches to a single branch.

For a given ground-truth image y from the target domain, the encoder EC^S of the self-representation network S encodes it into a latent space f^S,

$$f^S = EC^S(t) \tag{3}$$

Similar to the translation network, the latent features f^S are utilized to feed the decoder DC^S of the self-representation network S and extract the feature maps $f^{S,DC}$:

$$f_i^{S,DC} = DC^{S,i}(f^S) \tag{4}$$

where $DC^{S,i}(\cdot)$ denotes the forward computation process of the decoder in S.

The pseudo images $\hat{y} = T(x_{m_1}, x_{m_2}, x_{m_3})$ and $y^{'} = S(y)$ generated by the translation network and the self-representation network are all utilized to train the discriminator network D with the ground-truth image y.

Discriminator network Auto-GAN utilizes the network architecture of 70×70 "PatchGAN" (Isola et al. 2017; Zhu et al. 2017) in the discriminator network. Different from telling whether each pixel of input image is real or fake, this discriminator tries to classify whether each patch in the input image is real or fake. Such a patch-level discriminator penalizes structured errors at the scale of patches and has fewer parameters than a full image discriminator.

Modality mask vector Taking a modality mask vector as the target modality label, Auto-GAN can translate the images from multiple modalities to any missing modality with the same generator. The modality mask label is a matrix in the same size as the training images, and they are concatenated together as the inputs. The elements of each matrix in the modality mask vector shares the same value for each target modality label.

Network Losses

In Auto-GAN, we introduce three losses: self-representation loss, collaborative discriminator loss, and multiple generator loss.

Self-representation loss One of the main concepts of our proposed Auto-GAN is the self-representation loss. Different from traditional GAN-based image translation methods, Auto-GAN is supervised by not only *pixel-level* losses, but also *feature-level* losses. As aforementioned, the self-representation network S is trained by target image itself. When it is trained together with the translation network T, network S will better model the distribution of target images than the translation network T does. Therefore, we introduce the feature maps of decoder network DC^S to guide the decoder network DC^T at the feature level. Given four modalities, our proposed model can generate any missing modality from the other three modalities. The self-representation loss $\mathcal{L}^{SR}_{m_{k_4}}$ of generating modality m_{k_1} from $\{m_{k_2}, m_{k_3}, m_{k_4}\}$ can be defined as:

$$\begin{gathered}\mathcal{L}^{SR}_{m_{k_1}} = \sum_{i}^{n} \| f^{S,DC}_{i,m_{k_1}} - f^{G,DC}_{i;m_{k_2},m_{k_3},m_{k_4}} \|_2 \\ j, k_j \in \{1,2,3,4\}\end{gathered} \tag{5}$$

where $\| \cdot \|_2$ denotes the l_2-norm, $DC^i(\cdot)$ denotes the output of i-th layer in decoder network, and n denotes the number of convolutional layers in decoder networks DC^S and DC^T. In general, the self-representation loss $\mathcal{L}^{SR}$ can be written as:

$$\mathcal{L}^{SR} = \sum_{k}^{N} \mathcal{L}^{SR}_{m_k} \tag{6}$$

where $k \in \{1, 2, 3, \cdots, N\}$, and N denotes the number of modalities.

Collaborative discriminator loss The discriminator is utilized to predict whether an input image is real or fake. As aforementioned, the self-representation network S can estimate the distribution of target domain more accurately than translation network T does, we incorporate not only the pseudo image $\hat{y}$ translated by T but also the pseudo image $y^{'}$ generated by S to train the decoder network, together with the ground truth image S. Therefore, the collaborative discriminator loss $\mathcal{L}^{CD}$ can be computed as:

$$\begin{aligned}\mathcal{L}^{CD}(\hat{y}, y^{'}, y) = & \ \mathbb{E}_{y \sim P_y}[\log(D(y))] \\ & + \lambda_1 \cdot \mathbb{E}_{\hat{y} \sim P_{\hat{y}}}[\log(1 - D(\hat{y}))] \\ & + (1 - \lambda_1) \cdot \mathbb{E}_{y^{'} \sim P_{y^{'}}}[\log(1 - D(y^{'}))]\end{aligned} \tag{7}$$

where P_y, $P_{\hat{y}}$, and $P_{y'}$ are the distributions of the ground-truth image, the pseudo image translated by T, and the pseudo image generated by S, respectively. $\lambda_1 \in (0, 1)$ is a trade-off parameter between the self-representation network and the translation network.

Multiple generator loss As our model can generate any missing modality from the other three modalities, the generator loss is the sum of four different input combinations. To avoid the blurring effect (Isola et al. 2017) caused by l_2 loss (Pathak et al. 2016), we take the l_1 loss as the pixel-level loss to supervise the translation network T and self-representation network S. When m_1 is the target modality, the l_1 generator loss $\mathcal{L}^{MG,T}_{m_1}$ of translation network T and $\mathcal{L}^{MG,S}_{m_1}$ of self-representation network S can be computed as:

$$\begin{aligned}\mathcal{L}^{MG,T}_{m_1} &= \mathbb{E}_{x \sim P_x}[\| T(x_{m_1|m_2,m_3,m_4}) - y \|_1] \\ \mathcal{L}^{MG,S}_{m_1} &= \mathbb{E}_{x \sim P_x}[\| S(x_{m_1}) - y \|_1]\end{aligned} \tag{8}$$

Therefore, the multiple generator loss $\mathcal{L}^{MG,T}$ of translation network T and $\mathcal{L}^{MG,S}$ of self-representation network S can be written as:

$$\mathcal{L}^{MG,i} = \sum_{k}^{N} \mathcal{L}^{MG,i}_{m_k}, i \in \{T, S\} \tag{9}$$

Our full objective is:

$$\mathcal{L} = \mathcal{L}^{SR} + \mathcal{L}^{CD} + \lambda_2 \cdot \mathcal{L}^{MG,T} \tag{10}$$

$\mathcal{L}^{MG,S}$ is utilized to optimize the self-representation network S.

Experiments

To validate the effectiveness of the proposed Auto-GAN, we evaluate our method by two experiments representing different clinical scenarios: magnetic resonance (MR) image translation and CT image translation. We conduct the first experiment on the BraTS database (Menze et al. 2015), which consists of four MRI modalities: T1, T1-C, T2, and T2-FLAIR. Auto-GAN is used to generate any missing modality from the other three remaining modalities. For CT image translation, we evaluate our method on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database, which has two modalities: T1 and CT. Fig. 2 presents exemplary samples from these databases.

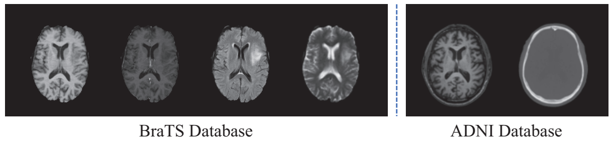

Figure 2: Exemplary examples of BraTS database. From left to right, they are T1, T1-C, T2-FLAIR, and T2; Exemplary examples of ADNI database, from left to right, they are T1 and CT.

Dataset and Protocols

MR image translation For MR image translation, we evaluate our method on the BraTS2018 dataset, which consists of 352 subjects with high- or low-grade gliomas. Each subject has four MRI modalities: T1, T1-C, T2, and T2-FLAIR. Since the tumor entity has different appearances in different modalities, the challenge for the translation task is much greater. The size of each MR image is $240 \times 240 \times 155$ with the voxel size of $1 \times 1 \times 1\ mm^3$.

CT image translation To validate the generalization ability of the proposed method, we extend the experiment from translating medical images among MRI modalities to MRI to CT translation. We evaluate CT image translation performance based on Auto-GAN on the ADNI database with 16 subjects. Each subject has a T1 image and a paired CT image. The MR images were scanned by a Siemens Trio TIM scanner, with the voxel size $1.2 \times 1.2 \times 1\ mm^3$, TE 2.95 ms, TR 2300 ms, and flip angle 9°. The voxel size of the corresponding CT images, which were scanned by a Siemens Somatom scanner, is $0.59 \times 0.59 \times 3\ mm^3$.

Dataset protocols We randomly select 80% subjects as a training set; the remaining 20% subjects are taken as a testing set. Such a process is repeated by 10 times. Unless explicitly mentioned, all the reported quantitative assessment is evaluated on the testing set. All the quality assessment for the perceptual appearance is based on the same protocol with the state-of-the-art methods.

Experimental Settings

We conduct all the experiments under the environment of Python 3.7 and PyTorch 1.0 on a Ubuntu 18.04 system with NVIDIA TITAN Xp GPU. All the images used in our experiments are spatially aligned. According to the slice-based scanning principle of medical images, the 3D medical images are only continuous on the scanning direction. The discontinuity on the other two directions make 3D convolution unsuitable. Therefore, based on the scanning direction, we cut each data into multiple slices and utilize 2D slices to train the proposed model. In the training phase, the input images are resized to 256×256 and then cropped to the size of 240×240. In the testing phase, the input images are not resized. The trade-off parameter λ_1 is set to 0.5 and λ_2 is set to 10 in our experiments.

To objectively assess the quantitative score of translated images, structural similarity index metric (SSIM) (Wang et al. 2004) and feature-similarity index (FSIM) (Zhang et al. 2011) as the evaluation criterion. All the real images from target modality are taken as reference dataset. The SSIM and FSIM scores of the translated images are taken as the quantitative evaluation.

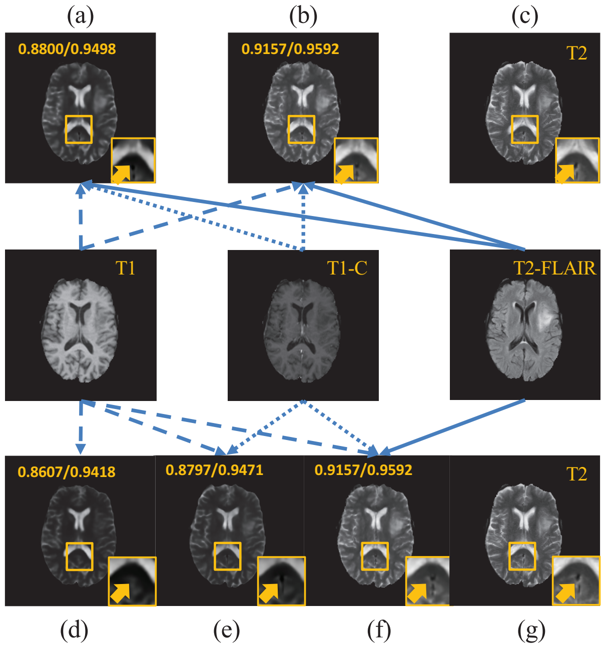

Figure 3: (a) is generated without self-representation constraint; (b) is generated with self-representation constraint; (d) is translated only from T1 modality; (e) is translated from both T1 and T1-C modalities; (f) is the translated from all three available modalities; (c) and (g) are the ground-truth T2 image. The yellow numbers are the SSIM scores and FSIM scores.

Ablation Study

We explore the improvements benefiting from the two key concepts in the proposed framework: self-supervised learning and collaborative learning.

To validate the effectiveness of self-supervised learning, we remove the self-representation network from the proposed framework and compare the experimental results with proposed approach on BraTS database, as shown in (a) and (b) of Fig. 3. {T1, T1-C, T2-FLAIR} are taken as inputs. Without self-supervised learning network, the synthesized result is indistinct with poor perceptual appearance and lost useful texture information (yellow arrow in (a) of Fig. 3). However, with self-supervised learning network, the synthesized results achieves much better perceptual appearance with more details. We highlighted the interhemispheric cerebral spinal fluid posterior to the splenial part of the corpus collosum, where (a) is not clear enough and the intensity distribution is not accurate. In (d)-(f), without collaborative learning, this structure is either completely lost or distorted. We also quantitatively assess the result by using SSIM and FSIM scores. The proposed method with self-supervised learning achieves 0.9157 in SSIM and 0.9592 in

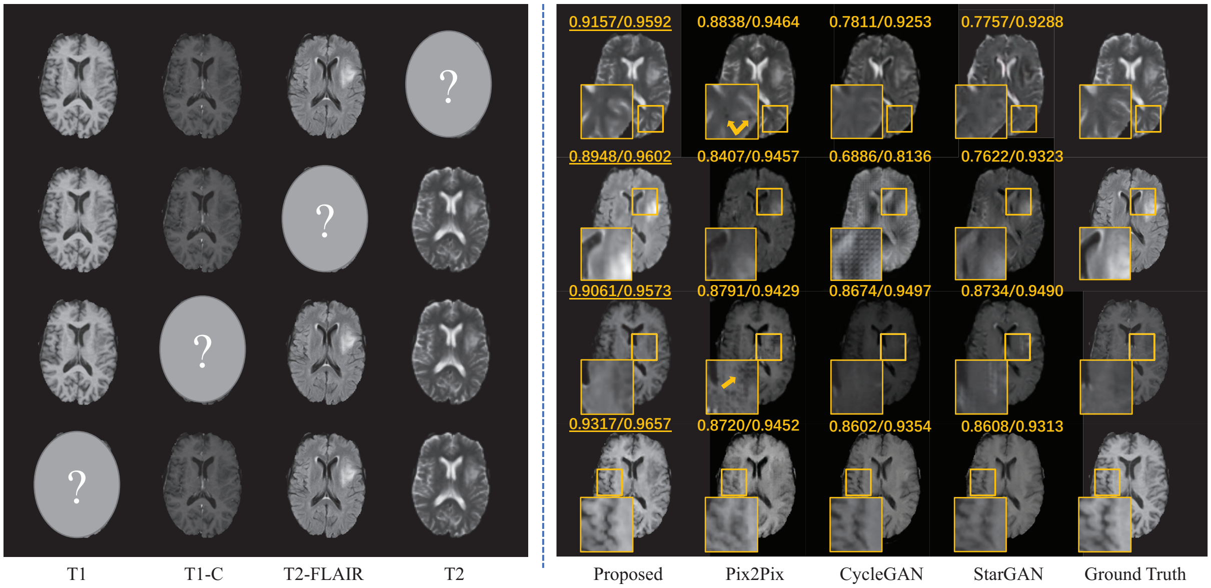

Figure 4: Experimental results of the proposed method, Pix2Pix, CycleGAN, and StarGAN on BraTS database. The yellow numbers are the SSIM and FSIM scores. The arrows point out the remarkable parts of the results. For Pix2Pix, CycleGAN, and StarGAN, synthesized T1/T1-C/T2 images are translated from T2-FLAIR, T2-FLAIR images are translated from T2. The results of Auto-GAN are generated from all the available modalities.

FSIM, much higher than 0.8800 and 0.9498 without self-representation learning.

To validate the effectiveness of collaborative learning, we change the branch number of the encoder in translation network from one to three. For ease of explanation, we take the translation of T2 modality as examples. In this experiment, the combinations of inputs are {T1}, {T1, T1-C}, and {T1, T1-C, T2-FLAIR}. As shown in (d), (e) and (f) of Fig. 3, (f) is very similar to the ground-truth image in (g). (e) shows better perceptual appearance than (d), indicating that more inputs lead to better results. Specifically, as aforementioned, when a subject has a lesion in the soft tissue, images from different modalities can provide unique information to enhance the synthesis results. For instance, T1 has clear texture for soft tissue but less tumor details. Therefore, the result in (d) contains less tumor lesion information. Because T1-C provides tumor parenchyma information and T2-FLAIR provides tumor edema information, as they were incorporated to train the model, the results show more accurate soft-tissue texture and better tumor lesion contrast. In addition, more inputs further increase the quantitative assessment scores.

Comparison

We compare our method with three popular and state-of-the-art GAN-based methods: Pix2Pix (Isola et al. 2017), CycleGAN (Zhu et al. 2017), and StarGAN (Choi et al. 2018).

MR image translation To evaluate the performance of the proposed method on multi-modality synthesis, the four modalities in the training set of the BraTS database are all utilized to train the Auto-GAN. Then, we generate each modality from the other three modalities using the testing set. As Pix2Pix, CycleGAN, and StarGAN are yield to a single input, we take T2-FLAIR as the input modality, because it provides more tumor lesion information than the other three modalities. T2-FLAIR modality is estimated by T2 modality, which is the most similar modality to T2-FLAIR.

As shown in Fig. 4, the synthesized results of the proposed method are very similar to the ground truth, while CycleGAN and StarGAN show poor perceptual appearance in generating T1-C, T2-FLAIR, and T2. Although Pix2Pix shows clear texture details, the results are noisy (as indicated by yellow arrows in Fig. 4) and the perceptual appearance is seriously degraded. All the methods show acceptable results for T1, but the tumor lesion is missing in the results of CycleGAN and StarGAN, and the details of the Pix2Pix's result are indistinct. The results of the proposed Auto-GAN show clear details of soft-tissue and distinct texture of tumor lesion area, which is superior to the other methods.

For the quantitative assessment, the SSIM and FSIM scores are shown on the top left of each sample, which are computed by the synthesized images and the corresponding ground truth. Due to the other GAN-based methods only utilize the pixel-level generator loss and a single input, they can not learn accurate distribution of the target modality in the feature level and miss the complementary information from multiple input modalities, which reduced the quantitative SSIM and FSIM scores. Our method achieves the highest quantitative scores on all the four modalities.

Note that, in these experiments, Pix2Pix, CycleGAN, and StarGAN need to be trained multiple times, each generating a different model. This is because they only allow to take a single fixed modality as an input. As a comparison, our pro-

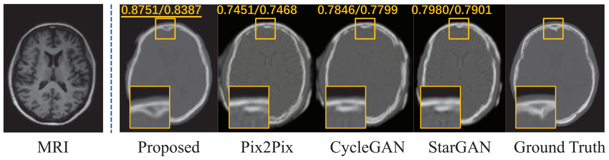

Figure 5: Experimental results of the proposed method, Pix2Pix, CycleGAN, and StarGAN on the ADNI database. Pseudo CT image are generated from the input MR image.

posed Auto-GAN can estimate any missing modality from the other available modalities in a unified, single model and achieves superior performance both qualitatively and quantitatively, which can be more efficient in the testing phase.

CT image translation As the existing GAN-based methods can only take one modality as input, for a fair comparison, we extend our framework to accommodate this scenario and decrease the branch number of the encoder in the translation network to one. The reduced framework thus only benefits from feature-level self-representation learning.

This experiment is conducted on the ADNI database, which consists of T1 and corresponding CT images. In clinical practice, the radiation effect caused by CT scan could potentially affect patient's health, while MR scan is considered safe and non-invasive. Therefore, in the experiments, we translate the images from T1 to CT modality. For qualitative evaluation, Pix2Pix, CycleGAN, and StarGAN show poor perceptual appearance with unexpected details and great deformation around the skull (yellow boxes in Fig. 5). The reduced framework of the proposed Auto-GAN achieves qualified pseudo CT images with more accurate and clearer skull contour. For quantitative evaluation, we compute the SSIM and FSIM scores from the synthesized CT images compared to the ground-truth CT images. As shown in the bottom of each sample, our method achieves the highest SSIM and FSIM scores, superior to the other competing methods. This experimental result further validates the effectiveness of the proposed feature-level self-supervised learning method.

For fair comparisons, we also use Inception Score (IS) (Salimans et al. 2016) and Fréchet Inception Distance (FID) (Heusel et al. 2017) as our evaluations. Our method achieves 2.15 of IS and 64.29 of FID. As a comparison, pix2pix, CycleGAN, and StarGAN achieve 1.21, 1.37, 1.05 of IS and 151.36, 129.61, 195.93 of FID, respectively.

Evaluation on Generalization Ability

To verify the generalization ability of the proposed method in broader applications, we conduct two additional experiments using natural image transformation tasks. We applied the same settings as those in the experiments of medical image generation on the CMP Facade database (Tyleček and Šára 2013) and the CUHK Student database (Wang and Tang 2008). As shown in Fig. 6, the proposed method can handle different translation tasks, indicating its strong generalization ability.

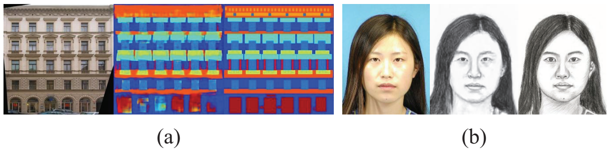

Figure 6: (a) is the experimental results on the CMP Facade database. From left to right, the examples are facade, generated labels and ground-truth labels. (b) is the experimental results on the CUHK Student database. From left to right, the examples are photo, generated sketch, and ground-truth sketch.

Limitations and Discussion

In this paper, we propose an Auto-GAN. Our method leverages the auto-encoder network to conduct self-supervised learning for a better guidance to each layer of the decoder in the domain translation network. Collaborative learning framework utilizes multi-facet information from multiple modalities. The designed modality mask vector empowers our Auto-GAN to generate any missing modality in a single, unified model, further guaranteeing its generalizability.

While our proposed Auto-GAN achieves superior performance compared to other state-of-the-arts, there are several limitations. For instance, in the training phase, more computing resources and computing time are required by the proposed framework. In the future, we will explore more efficient network architectures to handle more realistic and complex applications.

Acknowledgments

B.C., N.W., and X.G. were supported in part by the National Natural Science Foundation of China (61922066, 61876142, 61671339, 61772402, U1605252, and 61432014), the National Key Research and Development Program of China (2016QY01W0200 and 2018AAA0103202), the National High-Level Talents Special Support Program of China (CS31117200001), the Fundamental Research Funds for the Central Universities (JB190117), the Innovation Fund of Xidian University, and the Xidian University-Intellifusion Joint Innovation Laboratory of Artificial Intelligence.

References

Burgos, N.; Cardoso, M. J.; Thielemans, K.; Modat, M.; Pedemonte, S.; Dickson, J.; Barnes, A.; Ahmed, R.; Mahoney, C. J.; Schott, J. M.; et al. 2014. Attenuation correction synthesis for hybrid pet-mr scanners: application to brain studies. *IEEE transactions on medical imaging* 33(12):2332–2341.

Choi, Y.; Choi, M.; Kim, M.; Ha, J.-W.; Kim, S.; and 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*, 8789–8797.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Li, F.-F. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Gatys, L. A.; Ecker, A. S.; and Bethge, M. 2016. Image style transfer using convolutional neural networks. In *Proceedings of* *the IEEE conference on computer vision and pattern recognition*, 2414–2423.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Advances in Neural Information Processing Systems*, 6626–6637.

Hinton, G. E., and Salakhutdinov, R. R. 2006. Reducing the dimensionality of data with neural networks. *science* 313(5786):504–507.

Hinton, G.; Vinyals, O.; and Dean, J. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

Huang, P.; Li, D.; Jiao, Z.; Wei, D.; Li, G.; Zhang, H.; and Shen, D. 2019. Coca-gan: Common-feature-learning-based context-aware generative adversarial network for glioma grading. In *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention*.

Im, D. I. J.; Ahn, S.; Memisevic, R.; and Bengio, Y. 2017. Denoising criterion for variational auto-encoding framework. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Ioffe, S., and Szegedy, C. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, 448–456.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134.

Jog, A.; Carass, A.; Roy, S.; Pham, D. L.; and Prince, J. L. 2017. Random forest regression for magnetic resonance image synthesis. *Medical image analysis* 35:475–488.

Justin, J.; Alexandre, A.; and Li, F.-F. 2016. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, 694–711. Springer.

Kim, Y., and Rush, A. M. 2016. Sequence-level knowledge distillation. *arXiv preprint arXiv:1606.07947*.

Krizhevsky, A.; Sutskever, I.; and Hinton, G. E. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, 1097–1105.

Ledig, C.; Theis, L.; Huszár, F.; Caballero, J.; Cunningham, A.; Acosta, A.; Aitken, A.; Tejani, A.; Totz, J.; Wang, Z.; et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4681–4690.

Liu, Y.; Cao, J.; Li, B.; Yuan, C.; Hu, W.; Li, Y.; and Duan, Y. 2019. Knowledge distillation via instance relationship graph. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 7096–7104.

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

Menze, B. H.; Jakab, A.; Bauer, S.; Kalpathy-Cramer, J.; and Farahani, K. 2015. The multimodal brain tumor image segmentation benchmark (brats). *IEEE Transactions on Medical Imaging* 34(10):1993–2024.

Noh, H.; Hong, S.; and Han, B. 2015. Learning deconvolution network for semantic segmentation. In *IEEE International Conference on Computer Vision*, 1520–1528.

Pathak, D.; Krahenbuhl, P.; Donahue, J.; Darrell, T.; and Efros, A. A. 2016. Context encoders: Feature learning by inpainting. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2536–2544.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; Radford, A.; and Chen, X. 2016. Improved techniques for training gans. In *Advances in neural information processing systems*, 2234–2242.

Sauerbrei, W., and Royston, P. 1999. Building multivariable prognostic and diagnostic models: transformation of the predictors by using fractional polynomials. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 162(1):71–94.

Shen, D.; Wu, G.; and Suk, H.-I. 2017. Deep learning in medical image analysis. *Annual Review of Biomedical Engineering* 19(1):221–248.

Simonyan, K., and Zisserman, A. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Tyleček, R., and Šára, R. 2013. Spatial pattern templates for recognition of objects with regular structure. In *German Conference on Pattern Recognition*, 364–374. Springer.

Van Buuren, S.; Boshuizen, H. C.; and Knook, D. L. 1999. Multiple imputation of missing blood pressure covariates in survival analysis. *Statistics in medicine* 18(6):681–694.

Vemulapalli, R.; Nguyenb, H. V.; and Zhou, S. K. 2015. Unsupervised cross-modal synthesis of subject-specific scans. In *Proceedings of the IEEE International Conference on Computer Vision*, 630–638.

Wang, X., and Tang, X. 2008. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(11):1955–1967.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; Simoncelli, E. P.; et al. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13(4):600–612.

Yu, Y.; Tang, S.; Aizawa, K.; and Aizawa, A. 2018. Category-based deep cca for fine-grained venue discovery from multimodal data. *IEEE Transactions on Neural Networks and Learning Systems* 30(4):1250–1258.

Yu, Y.; Tang, S.; Raposo, F.; and Chen, L. 2019. Deep cross-modal correlation learning for audio and lyrics in music retrieval. *ACM Trans. Multimedia Comput. Commun. Appl.* 15(1):20:1–20:16.

Zhang, L.; Zhang, L.; Mou, X.; and Zhang, D. 2011. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing* 20(8):2378–2386.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, 2223–2232.