Joint Attention GAN for Multimodal Image Translation

***Abstract*—Multimodal images are required in many practical scenarios, ranging from clinical diagnosis to public security. Data missing often leads to decision bias. Although image translation achieved great progress, multimodal image translation is still challenging due to the difficulty in modeling correlations between multiple inputs. In this paper, we novelly proposed a joint attention GAN (MA-GAN) framework to generate the missing image modality through multimodal available images. To effectively extract the multimodal representation specific to desired modality, we use a self-representation network to drive a inter-modal attention module. The extracted multimodal features can maintain kernel-level consistency with the target modality, which greatly improved the image translation performance. Quantitative and qualitative comparisons in various multimodal image generation tasks against several prior methods demonstrate the superiority of our approach, by showing more precise and realistic results.**

***Index Terms*—Multimodal, GANs, Attention**

I. Introduction

In many applications of image processing, computer vision, and computer graphics, multimodal image data is required but only partially available due to various practical reasons. For example, modern medicine, especially precision medicine, has increasingly depended on multimodal medical images (*e.g.,* , native (T1), post-contrast T1-weighted (T1Gd), T2-weighted (T2) and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR)) [5]. Each modality can provide certain *in vivo* contrast and resolution in visualization of different internal human organs. While doctors desire to joint multiple insights from the complete set of multimodal medical images to make a more precise diagnostic decision, unfortunately, certain practical limitations (*e.g.,* , restricted medical conditions, inadequate scanning time, and cost/spending/resource control) could result in imaging modal with systematic errors or even sacrificing some imaging modals [12]. These unavailable images can cause decision bias for doctors and statistical techniques.

Since the restrictions on acquiring the missing or low-quality images in practical, imputing the unavailable modalities by using the remaining clean image data has become an important research topic in computer vision and image processing. The estimated missing values can provide supplementary information to doctors, besides the available image data.

Multimodal image data imputation can be regarded as an image translation issue. Similar tasks include style transfer, segmentation, deblurring, super-resolution, etc, which translates the images from one domain to another. Different domains refer to various imaging differences among light conditions, facial expressions, sensors, etc. Researchers have devoted considerable effort to developing effective image translation algorithms, and achieved significant progress on generative adversarial networks (GANs) [6] based methods, such as cGAN [7], CycleGAN [20] and StarGAN [3]. Unfortunately, these approaches are only qualified for single modality input, but the valid information contained in other available modalities is not taken into account. CollaGAN [10] is one of the few works designed for multiple inputs. But the method simply parallels multimodal image inputs without effective constraints on them, especially the target-modal-specific supervision.

To effectively extract the multimodal representation specific to desired modality, we propose modal complementing generative adversarial networks (MA-GAN) guided by a self-representation network for multimodal image translation. We follow the hourglass network [7] architecture mostly used in image translation tasks, the proposed framework consists of translation network and self-representation network. Our MA-GAN is constrained in two aspects: *inter-modality complementarity* and *cross-modality consistency*. For complementarity, a modal complementing (MC) module is designed to supplement current input modality with information from other input modalities. This supplement mechanism further requires feature compatibility between modalities. Therefore, we introduce a self-representation network during training to achieve cross-modality and kernel-level consistency. On the other hand, the cross-modality consistency constraint guarantees that only features consistent to target modality will be extracted and irrelevant information will be filtered out.

In summary, our contributions are listed as follows:

- We propose modal complementing module to supplement each modality with information from other input modalities.
- We for the first time introduce self-representation network to guide the generative model during training. This provides feature compatibility demanded by modal complementing module and filters out irrelevant information in feature extraction stage.
- Quantitative and qualitative comparisons in these multimodal image generation tasks against several prior methods demonstrate the superiority of our approach, by showing more precise and realistic results.

II. Related Work

Existing image translation methods can be categorized into data-driven methods and model-driven methods.

Data-driven methods [2], [14], [16], [19] translated images according to the samples in the training set. All the images are cropped into patches. These methods learn a set of reconstruction coefficients for the patches of a testing image. Then, the target image is synthesized by reconstruction coefficient

and image patches in the training set. For example, Wang and Tang [16] employed a multi-scale Markov random field model to preserve the local information for face image-to-image translation, which takes one candidate patch for each testing patch. This method is further improved by Zhou *et al.* [19], which increases the number of candidates. These methods heavily depend on the integrity of training data. Accurate translated results require large training set, which increases computational complexity. Meanwhile, the patch-based synthesized results are blurring effects with poor perceptual appearance.

Model-driven methods [3], [7], [10], [18], [20] learned a mapping function to translate the image from source to target domain directly. Thanks to generative adversarial networks (GANs) [6], these methods achieved significant improvement on generating realistic images. For image translation tasks, GANs are further improved by imposing an image condition as input, which is called conditional generative adversarial networks (cGAN). Specifically, Pix2Pix [7] learns to translate the input image condition to the fake sample. It requires paired training data. In practice, however, paired data is hard to acquire, and the unpaired data can cause Pix2Pix to produce great deformation. To extend Pix2Pix to unpaired data, CycleGAN [20] and DiscoGAN [9] applied a cycle-consistent loss to preserve the image core aspects by reconstructing an input image from a generated fake sample. However, Pix2Pix and CycleGAN are designed for translating images between two different domains. They are not qualified for multimodal image translation tasks. Through adding target modality mask vector to the input of generative network, StarGAN [3] [4] successfully translated an image from one domain to multiple different domains. Here, the discriminator is also improved to multi-task architecture, which outputs not only the real/fake label but also the class label. However, similar to Pix2Pix and CycleGAN, StarGAN still cannot deal with multimodal data inputs, which makes it impossible to effectively utilize all of the remaining data.

CollaGAN [10] is a recently proposed framework to process multimodal data inputs, which redefined a multiple cycle consistency loss for multimodal image translation tasks. The discriminator is similar to the one of StarGAN. An SSIM loss is utilized to further improve the generative performance. However, the relationships among multimodal image inputs are not exploited.

III. METHOD

To better extract the multimodal representation specific to desired modality, we propose a novel framework for multimodal image translation, which joint attention and self-supervised learning in a unified manner. The proposed MA-GAN can generate any missing modality from available modalities flexibly.

A. Motivation and overview

Existing GAN-based image translation methods achieved great progress in generating realistic images between two different domains. However, few of these methods could take multimodal images as the inputs. The most recent work on multimodal image translation task is CollaGAN [10], which simply concatenates multimodal images together in pixel-level or feature-level. The complementary information among all the available multimodal images are ignored in these methods.

Here, we attempt to fuse inter-modality complementary information to improve the performance of multimodal image translation task. However, most multimodal image translation tasks lack of large scale datasets, which makes it hard to furthest take advantage of inter-modality complementarity. To tackle this problem, we propose modal complementing module to supplement each modality features with features extracted from other input modalities.

Inspired by [17], we propose inter-modality complementarity, inter-modality compatibility, and cross-modality consistency for multi-modal image translation task and incorporate them in a unified generative framework(MA-GAN):

- Inter-modality complementarity: For multimodal tasks, most modality contains information that other modalities does not. This nature can be utilized to impute better missing modality.
- Inter-modality compatibility: This is the premise of inter-modality complementarity, which guarantees that there will be no loss of information when taking advantage of complementarity.
- Cross-modality consistency: Only features that is consistent(relevant) to target modality must be extracted and others should be filtered out.

Our MA-GAN consists of self-representation network and translation network with embedded modal complementing module, as shown in Fig. 1. Out of inter-modality complementarity, modal attention adds information to each input modality from other modalities before feature fusion. Considering the strong self-representation ability of auto-encoder, we hereby borrow auto-encoder network as the self-representation network to guide the generaotr to provide inter-modality compatibility for modal attention module, *i.e.,* additivity between features brought by kernel-level supervision. Cross-modality consistency is also implemented by supervision of self-representation network.

Without loss of generality and for ease of representation, we assume there are four modalities $\{m_1, m_2, m_3, m_4\}$ in the datasets. For translation network, we assume that the input images $\{x_1, x_2, x_3\}$ in modalities $\{m_1, m_2, m_3\}$ are translated to the target image x_4 in target modality m_4.

B. Modal complementing module

Each channel in the feature maps can be regarded as a certain pattern extracted from the input by the deep neural network. For the k-th channel (or pattern), the information contained in each input modality is different. Taking the medical images as an example, if the k-th channel is about the texture feature of the tumor, obviously T1Gd contains the most abundant and valuable information; if the channel is about the cerebrospinal fluid, then T2 is the most important modality. A natural idea is to add the tumor-related information in

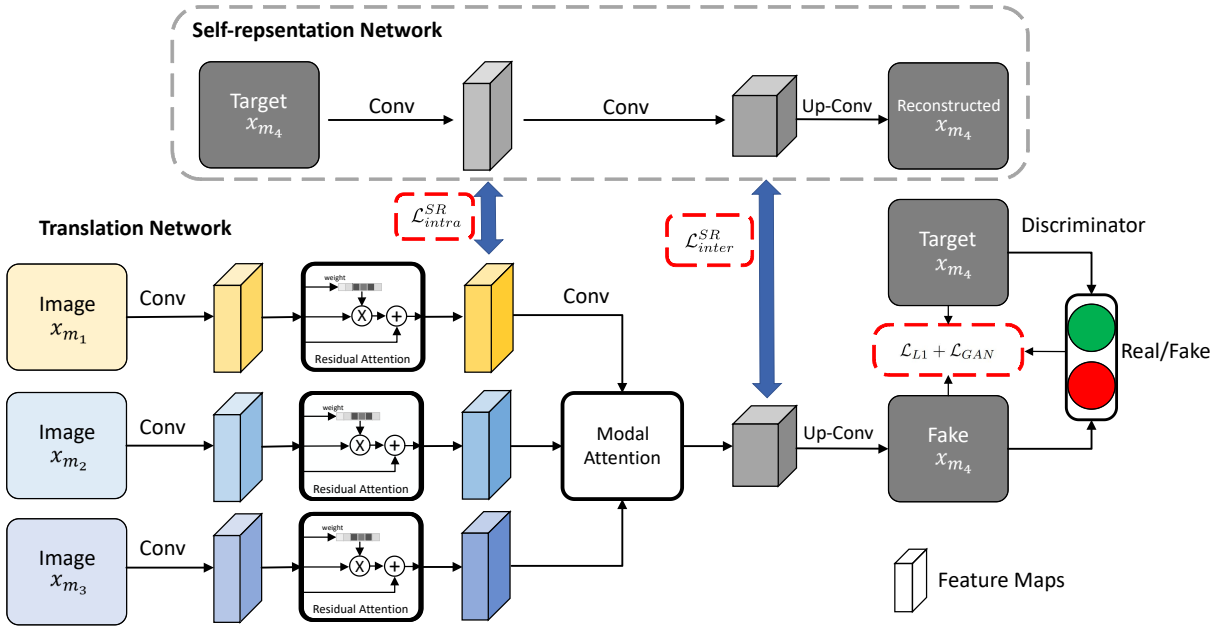

Fig. 1. Joint Attention GAN framework overview. Input images $\{x_{m_1}, x_{m_2}, x_{m_3}\}$ from modalities $\{m_1, m_2, m_3\}$ are fed to the corresponding branch of our translation network to generate target image x_{m_4} in missing modality m_4. The feature maps are extracted by combining MA module and RA module, supervised by self-representation network.

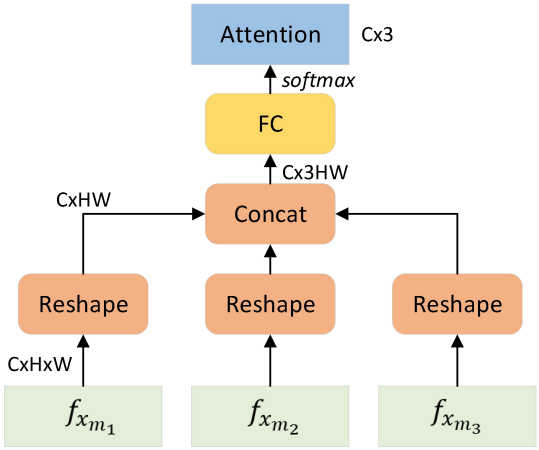

Fig. 2. Attention matrix calculation in inter-modal attention. FC denotes two fully connected layers defined in Equation 1

T1Gd to T2 modality, and add the cerebrospinal fluid-related information in T2 to T1Gd modality.

Note that, before supplementing information to other modality, for a given channel or pattern, we need to know which modality contains the best information or which modality we need to pay more attention to. Motivated by attention mechanism [1] [13] [15] [18], we propose a modal attention to calculate the weights (denotes as $\mathcal{A}$) of cross-modality channels, as shown in Fig. 2,

$$\mathcal{A} = S_2(\sigma((\delta([f_{x_1}, f_{x_2}, f_{x_3}]W_1))W_2)) \tag{1}$$

where $f_{x_k} = E_k(x_k)$, $E_k(\cdot)$ denotes the output of k-th encoding branch E_k of generator G, $[\cdot,\cdot]$ denotes feature reshaping (flatten a tensor from CxHxW to CxHW) and concatenation. W_1 and W_2 are two fully connected layers, as shown in Fig. 2. δ and σ are ReLU [11] and Sigmoid function, respectively. S_2 is softmax function along each row of $\mathcal{A}$. In other words, the scalars in each row of $\mathcal{A}$ are the weights of all modalities to a certain pattern, and the scalars in each column of $\mathcal{A}$ are the weights of a certain modality to all patterns.

As shown in Figure 3, we use $\mathcal{A}$ to supplement the information of other modalities to the input image. In this way, we can get the complementary feature $f_{x_i}^{comp}$ specific to the i-th modality:

$$f_{x_i}^{comp} = \gamma * f_{x_i} + (1-\gamma) * \sum_{k=1}^{n} (f_{x_k} * \mathcal{A}_k) \tag{2}$$

where $*$ denotes the channel-wise multiplication between a vector and the feature maps, n is the number of input modalities *i.e.,* 3 under our assumption. $\mathcal{A}_k$ is the k-th column of attention weight matrix $\mathcal{A}$. While information supplementing, the information of current modality is preserved by setting a trade-off parameter γ.

C. Self-supervised learning

As we aforementioned, an auto-encoder is introduced to drive the translation network $\mathcal{T}$, which consists of a generator $\mathcal{G}$ with multi-branch encoder and a single branch decoder, and discriminator $\mathcal{D}$. The auto-encoder is taken as self-representation network SR.

SR drive $\mathcal{T}$ by kernel-level supervision:

$$\mathcal{L}_{SR} = \sum_{k=1}^{n} \|f_{x_i} - E(x_{m_4})\|_2 \tag{3}$$

where $\mathcal{G}^{SR,j_2}(\cdot)$ denotes the output of j_2-th layer in self-representation network $\mathcal{SR}$.

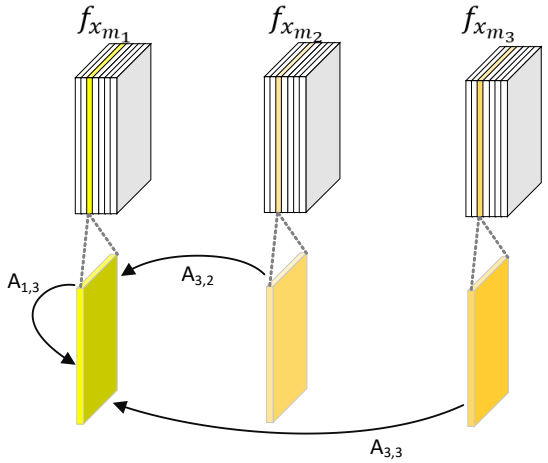

Fig. 3. Information supplementation. $A_{i,j}$ is the ith row jth column item of A. The first modality supplements the information of the 2nd mode on the 3th channel with weight $A_{3,2}$, and supplements the information of the 3th mode on the 3th channel with by weight $A_{3,3}$.

D. Network Loss

In the training phase, we use Adam to optimize the total loss of generator G of MA-GAN:

$$\mathcal{L}_G = \mathcal{L}_{L1} + \mathcal{L}_{GAN} + \alpha \cdot \mathcal{L}_{intra}^{SR} + \beta \cdot \mathcal{L}_{inter}^{SR} \tag{4}$$

where $\mathcal{L}_{L1}$ and $\mathcal{L}_{GAN}$ are both defined in pix2pix, α and β are weights of $\mathcal{L}_{intra}^{SR}$ and $\mathcal{L}_{inter}^{SR}$ respectively.

E. Network Implementation

MA-GAN is composed of translation network and self-representation network.

Translation Network Translation Network consists of a generator and a discriminator. We adapt 70 × 70 PatchGANs [7] as our discriminator network and architecture from [8] for our generator. This network contains three convolutions for feature extraction and downsampling, several residual blocks for feature transformation, and three convolutions for upsampling. In order to adjust to multi-modal input, we duplicate the first three convolutions $\{Conv_1, Conv_2, Conv_3\}$ n times to form a multi-branch encoder where n is the number of input modalities and $Conv_k$ denotes the k-th convolution. We add RA attention module to $Conv_2$ and $Conv_2$, and add MA attention to $Conv_3$.To enable the model to receive any three modes as input, we add mask vector of target modality to every input image following [3] [10].

Self-representation Network The self-representation network adopts the same network structure as the generator, except that $\mathcal{L}_{GAN}$ and multi-branch encoder are not used. We feed original images into self-representation network without any mask. The experiment shows that only using $\mathcal{L}_{L1}$ supervision, the self-representation network can quickly fit on each dataset and effectively assist translation in the feature level.

IV. Conclusion

The conclusion goes here.

Appendix A
Proof of the First Zonklar Equation

Appendix one text goes here.

Appendix B

Appendix two text goes here.

Acknowledgment

The authors would like to thank...

References

[1] D. Britz, A. Goldie, T. Luong, and Q. Le. Massive Exploration of Neural Machine Translation Architectures. *ArXiv e-prints*, Mar. 2017.
[2] N. Burgos, M. J. Cardoso, K. Thielemans, M. Modat, S. Pedemonte, J. Dickson, A. Barnes, R. Ahmed, C. J. Mahoney, J. M. Schott, et al. Attenuation correction synthesis for hybrid pet-mr scanners: application to brain studies. *IEEE transactions on medical imaging*, 33(12):2332–2341, 2014.
[3] Y. Choi, M. Choi, M. Kim, J.-W. Ha, S. Kim, and J. Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018.
[4] Y. Choi, Y. Uh, J. Yoo, and J.-W. Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8188–8197, 2020.
[5] A. Drevelegas and N. Papanikolaou. Imaging modalities in brain tumors. In *Imaging of brain tumors with histological correlations*, pages 13–33. Springer, 2011.
[6] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
[7] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017.
[8] J. Johnson, A. Alahi, and L. Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016.
[9] T. Kim, M. Cha, H. Kim, J. K. Lee, and J. Kim. Learning to discover cross-domain relations with generative adversarial networks. *arXiv preprint arXiv:1703.05192*, 2017.
[10] D. Lee, J. Kim, W.-J. Moon, and J. C. Ye. Collagan: Collaborative gan for missing image data imputation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2487–2496, 2019.
[11] V. Nair and G. E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML*, 2010.
[12] L. N. Tanenbaum, A. J. Tsiouris, A. N. Johnson, T. P. Naidich, M. C. DeLano, E. R. Melhem, P. Quarterman, S. Parameswaran, A. Shankaranarayanan, M. Goyen, et al. Synthetic mri for clinical neuroimaging: results of the magnetic resonance image compilation (magic) prospective, multicenter, multireader trial. *American Journal of Neuroradiology*, 38(6):1103–1110, 2017.
[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.
[14] N. Wang, X. Gao, L. Sun, and J. Li. Bayesian face sketch synthesis. *IEEE Transactions on Image Processing*, 26(3):1264–1274, 2017.
[15] X. Wang, R. Girshick, A. Gupta, and K. He. Non-local neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7794–7803, 2018.
[16] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE transactions on pattern analysis and machine intelligence*, 31(11):1955–1967, 2008.
[17] C. Zhang, Y. Cui, Z. Han, J. Zhou, H. Fu, and Q. Hu. Deep partial multi-view learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.
[18] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. In *International Conference on Machine Learning*, pages 7354–7363. PMLR, 2019.

[19] H. Zhou, Z. Kuang, and K.-Y. K. Wong. Markov weight fields for face sketch synthesis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1091–1097. IEEE, 2012.
[20] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

PLACE PHOTO HERE

Michael Shell Biography text here.

John Doe Biography text here.

Jane Doe Biography text here.