UNSUPERVISED HIERARCHICAL TRANSLATION-BASED MODEL FOR MULTI-MODAL MEDICAL IMAGE REGISTRATION

*Xinru Dai, Tai Ma, Haibin Cai, Ying Wen**

Shanghai Key Laboratory of Multidimensional Information Processing,
School of Communication and Electronic Engineering, East China Normal University, Shanghai, China.

ABSTRACT

Deformable registration of multi-modal medical images is a challenging task in medical image processing due to the differences in both appearance and structure. We propose an unsupervised hierarchical translation-based model to perform a coarse to fine registration of multi-modal medical images. The proposed model consists of three parts: a coarse registration network, a modal translation network and a fine registration network. First, the coarse registration network learns to obtain the coarse deformation field, which is applied as structure-preserving information to generate a translated image by the modal translation network. Then, the translated image as enhancing information combined with the original images are used to derive a fine deformation field in the fine registration network. Furthermore, the final deformation field is composed from the coarse and the fine deformation fields. In this way, the proposed model can learn high accurate deformation field to implement multi-modal medical image registration. Experiments on two multi-modal brain image datasets demonstrate the effectiveness of this model.

Index Terms— Medical image registration, multi-modal medical image, modal translation, deep learning

1. INTRODUCTION

Multi-modal medical image registration is a prerequisite for the clinical diagnoses, and its difficulty lies in the differences in both appearance and structure between two images to be aligned. Traditional registration methods such as SyN [1] aligns the anatomical structures of medical images by using Mutual Information (MI) [2] as a driving force. However, traditional methods are time-consuming and not suitable for clinical applications. Popular unsupervised learning-based deformable image registration methods [3] [4] [5] [6] generally learn the deformation field by minimizing the intensity dependent loss functions such as Mean Square Error (MSE). Since the multi-modal images have large differences in appearance, MSE loss function is inappropriate. Some intensity independent loss functions such as Normalized Cross-Correlation (NCC) [7] and Negative Mutual Information (NMI) [8] are commonly used to obtain accurate multi-modal registration results. Meanwhile, Generative Adversarial Network (GAN) is applied in multi-modal medical image translation [9] [10] [11] [12] to reduce the differences in appearance of images. However, generative adversarial methods demand a large amount of training data and have difficulty in maintaining important anatomical structures for unpaired multi-modal images, generating unrealistic results.

In this paper, an unsupervised hierarchical translation-based network is proposed for multi-modal medical image registration. The model is integrated by three parts: a coarse registration network, a modal translation network and a fine registration network. The coarse deformation field obtained from the coarse registration network is used to perform the structure-preserving appearance translation in the modal translation network. Then the translation result is applied as enhancing information in the fine registration network to derive a fine deformation field. The coarse and the fine deformation fields are composed to the final deformation field, delivering accurate image registration results.

2. METHOD

We propose an unsupervised hierarchical translation-based model for multi-modal medical image registration. Overall structure of the proposed method is shown in Fig.1. The model contains three networks of U-Net [13] structure. Given multi-modal moving and fixed images, here a moving T2 image (M_{T2}) and a fixed T1 image (F_{T1}) are used. The proposed model outputs a deformation field ϕ, aiming to perform an accurate multi-modal medical image registration. Images are warped via Spatial Transformation Network (STN) [14].

2.1. Coarse Registration Network

Multi-modal images differ in both appearance and structure. The translation of intensity may change the anatomical struc-

This work was supported in part by 2030 National Key Research and Development Program of China (2018AAA0100500), the Natural Science Foundation of Shanghai (22ZR1421000), the National Nature Science Foundation of China (no.61773166), Projects of International Cooperation of Shanghai Municipal Science and Technology Committee (14DZ2260800) and the Fundamental Research Funds for the Central Universities. *Corresponding author: ywen@cs.ecnu.edu.cn

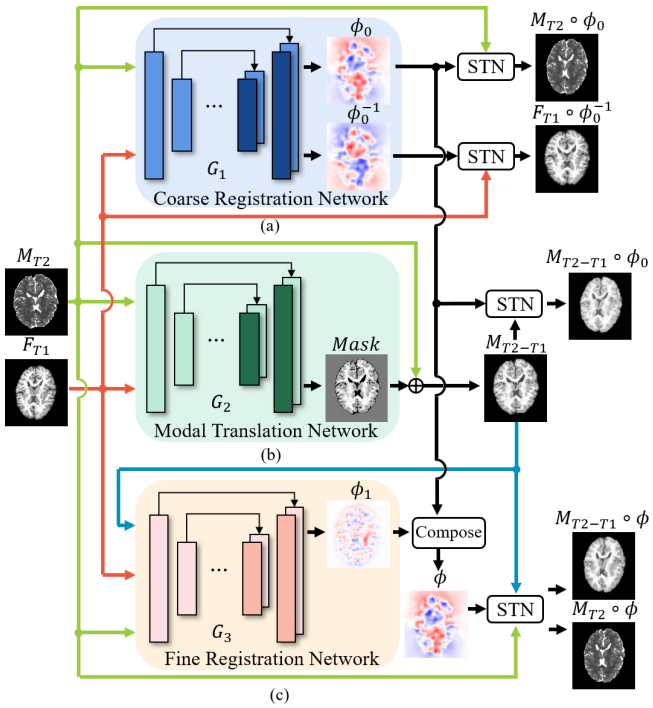

Fig. 1. An overview of the proposed multi-modal image registration model. (a), (b) and (c) represent the three networks of the proposed model. Lines in green, red and blue indicate different inputs. Lines in black denote the process of the networks.

tures of unpaired images with a large deformation. Therefore we propose a coarse registration network G_1 to keep the structure of images in the next modal translation network. The purpose of G_1 is to provide an initial deformation field for the following modal translation network and the fine registration network. As shown in Fig. 1 (a), taking F_{T1} and M_{T2} from different modalities as inputs. Images are defined on a n-D spatial domain $\Omega \subset \mathbb{R}^n$. G_1 searches for a coarse deformation field ϕ_0 and the corresponding inverse ϕ_0^{-1} to align the images: $\phi_0 = G_{1\theta_1}(F_{T1}, M_{T2})$, where θ_1 denotes the learnable parameters of G_1. There are significant appearance differences between F_{T1} and M_{T2}, thus the NMI loss function is chosen to penalizes the differences in structure:

$$\mathcal{L}_{NMI}(M_{T2} \circ \phi_0, F_{T1}) = -\sum_{a,b} p(a,b) \log \frac{p(a,b)}{p(a)p(b)}$$

where $p(a,b)$, $p(a)$ and $p(b)$ represent the joint and independent intensity distribution of two images. And $M_{T2} \circ \phi_0$ denotes the warped M_{T2} through the deformation field ϕ_0.

ϕ_0 is characterized by a displacement vector field u. A smooth deformation field is ensured by a regular term, which is employed on the spatial gradients of u:

$$\mathcal{L}_{smooth}(\phi_0) = \sum_{p \in \Omega} \|\nabla u(p)\|^2$$

We apply a bidirectional registration strategy to improve the accuracy and smoothness of the deformation field. Full objective of G_1 is shown as follows:

$$\begin{aligned}\mathcal{L}_1 = \alpha_1 \mathcal{L}_{smooth}(\phi_0) + \beta_1 \mathcal{L}_{NMI}(M_{T2} \circ \phi_0, F_{T1}) \\ + \beta_1 \mathcal{L}_{NMI}(F_{T1} \circ \phi_0^{-1}, M_{T2})\end{aligned} \tag{1}$$

where α_1 and β_1 represent the hyperparameters of G_1. The bidirectional NMI losses adopt the same hyperparameter [3].

2.2. Modal Translation Network

Due to the large differences in appearance of multi-modal images, modal translation is essential before the registration. With the help of the coarse registration network G_1, the modal translation network G_2 can generate an additive translation mask for structure-preserving appearance translation.

For F_{T1} and M_{T2}, G_2 learns to produce a translation mask: $Mask = G_{2\theta_2}(F_{T1}, M_{T2})$, where θ_2 denotes the learnable parameters of G_2. Then the $Mask$ is element-wise added to M_{T2}, generating a corresponding translated T1 image M_{T2-T1}. To retain the original anatomical structures of the moving image M_{T2} during the appearance translation, the NMI loss is used to penalize the structural difference between $Mask$ and M_{T2}. Moreover, as can be seen in Fig. 1 (b), the coarse deformation field ϕ_0 of G_1 is applied to warp M_{T2-T1}. Under this premise, M_{T2-T1} keeps its structure approaching M_{T2}, while $M_{T2-T1} \circ \phi_0$ has the same appearance and structure with F_{T1}. Thus the MSE loss can be used to optimize the network:

$$\mathcal{L}_{MSE}(M_{T2-T1} \circ \phi_0, F_{T1}) = \|M_{T2-T1} \circ \phi_0 - F_{T1}\|^2$$

Meanwhile the NMI loss is adopted to penalize the differences in structure of M_{T2-T1} and M_{T2}. The complete loss function proposed in this part is:

$$\begin{aligned}\mathcal{L}_2 = \alpha_2 \mathcal{L}_{NMI}(Mask, M_{T2}) + \beta_2 \mathcal{L}_{NMI}(M_{T2-T1}, M_{T2}) \\ + \lambda_2 \mathcal{L}_{MSE}(M_{T2-T1} \circ \phi_0, F_{T1})\end{aligned} \tag{2}$$

where α_2, β_2 and λ_2 represent the hyperparameters of G_2. As a consequence, the $Mask$ generated from G_2 can change the appearance of M_{T2} to be similar with that of F_{T1}, while keeping the anatomical structures of M_{T2}.

2.3. Fine Registration Network

The deformation filed ϕ_0 from the coarse registration network G_1 needs to be refined. The purpose of the fine registration network G_3 is to perform an accurate registration of the multi-modal images by using the modal translation results M_{T2-T1} from the modal translation network G_2.

The structure of the fine registration network is illustrated in Fig. 1 (c). To obtain the the fine deformation field, G_3 takes

F_{T1}, M_{T2} and M_{T2-T1} as the subjects of research: $\phi_1 = G_{3\theta_3}(F_{T1}, M_{T2}, M_{T2-T1})$, where θ_3 denotes the learnable parameters of G_3. M_{T2-T1} is applied as enhancing information, which is considered to have the same modality as F_{T1}. The final deformation field is derived from composing the coarse deformation field ϕ_0 and the fine deformation field ϕ_1: $\phi = \phi_0 \circ \phi_1$. Eq. (3) shows the full objective of G_3. The smooth loss is used to penalize ϕ_1, meanwhile the NMI and the MSE loss are applied to penalize ϕ:

$$\begin{aligned}\mathcal{L}_3 = \alpha_3 \mathcal{L}_{smooth}(\phi_1) + \beta_3 \mathcal{L}_{NMI}(M_{T2} \circ \phi, F_{T1}) \\ + \lambda_3 \mathcal{L}_{MSE}(M_{T2-T1} \circ \phi, F_{T1})\end{aligned} \tag{3}$$

where α_3, β_3 and λ_3 represent the hyperparameters of G_3.

3. EXPERIMENTS

3.1. Dataset and Preprocessing

The experiments of the proposed unsupervised hierarchical translation-based model focus on the application of T2-to-T1 2D multi-modal image registration. Two multi-modal brain image datasets are used to evaluate the method. The first one is the Brainweb [15], in which the structural differences between objects are not large enough. So we slice the 3D data, adding random shift, rotation, scaling and deformation to the images. Then we normalize and crop the images to the size of 128×128. The 492 generated images for each modal are divided into two groups for training (396 cases) and testing (96 cases). The second one is the Brats [16]. We slice the 3D data, normalizing and cropping the images to get 1710 cases for each modal with size of 176×176, in which 1416 are used for training and 294 for testing. In the experiments, we train the proposed model for subject-to-subject registration. The experiments are based on Keras with a Tensorflow [17] backend and the ADAM [18] optimizer, which are implemented on Intel(R) Xeon(R) Gold 6230 CPU.

The proposed networks G_1, G_2 and G_3 adopt the similar U-Net [13] architecture. Networks are trained by stages. Weights of each network is frozen on completion of the default 150,000 iterations training. We find the optimal hyperparameters by experiments, setting α_1 = 0.02, β_1 = 0.15 in Eq. (1); α_2 = 0.02 β_2 = 0.02, λ_2 = 2.5 in Eq. (2); α_3 = 0.02, β_3 = 0.1, λ_3 = 1.5 in Eq. (3).

The evaluation indexes applied in experiments are Dice score, Peak Signal to Noise Ratio (PSNR), Structural Similarity (SSIM) and MSE. Dice is used to measure the accuracy of the registration results on the Brainweb because of its complete label. Since the Brats only has the tumor label, which is individually special, we use MSE as index in evaluation. Specifically, the deformation field learned from the unpaired multi-modal images is applied to the unpaired uni-modal images, then the MSE loss is computed between these images of the same modality. PSNR and SSIM are adopted to evaluate the appearance and structure of the translation results.

3.2. Experiments of Registration

We compare the proposed model with Affine registration, SyN [1] implemented by ANTs [19] package and VoxelMorph (VM) [3]. Moreover, we extend VM with three similarity metrics: NMI, NCC and LG [20] for comparison, which is indicated in the parentheses. The optimal hyperparameters for each loss function are found by experiments. Results on the Brainweb are illustrated in Table 1. The Avg.Dice represents the average value of Dice score in segmented anatomical structures. Results on the Brats are shown in Table 2.

Table 1. Registration results on the Brainweb.

Method	Avg.Dice (%)	Runtime (s)
Affine	46.452	-
SyN [1]	57.506	2.260
VM [3] (LG)	54.670	0.049
Ours (LG)	56.599	0.052
VM [3] (NCC)	58.196	0.052
Ours (NCC)	61.702	0.058
VM [3] (NMI)	67.075	0.050
Ours (NMI)	**70.569**	0.059

Table 2. Registration results on the Brats.

Method	MSE (%)	Method	MSE (%)
VM [3] (LG)	2.825	Ours (LG)	2.650
VM [3] (NCC)	2.688	Ours (NCC)	2.525
VM [3] (NMI)	2.315	Ours (NMI)	**2.162**

Results from Table 1 and Table 2 indicate that the proposed model achieves a significant improvement with the addition of G_2 and G_3 networks compared to VM. Among three different similarity metrics, the proposed model using NMI loss obtains the most accurate result. Quantitatively, the Avg.Dice of the proposed model is generally 3% higher than that of VM. Besides, compared with SyN, our model remains an advantage in terms of time.

Fig. 2 and Fig. 3 are the visualizations of the registration results on the Brainweb and the Brats. For the Brats, the deformation field learned from M_{T2} and F_{T1} is applied on the corresponding M_{T1}. The area marked with the red box reveals that SyN and VM roughly align the contours of images, while our model achieves more accurate registration results for the internal anatomy in both datasets. As is illustrated in Fig. 2, the green and red masks represent the anatomical structures of the warped moving images and the fixed image. The overlap of masks reflects the accuracy of registration. Among all methods, the proposed model using NMI loss shows the best performance.

3.3. Experiments of Modal Translation

CycleGAN [21] is widely used in modal translation, thus we adopt it as the comparison for modal translation on the

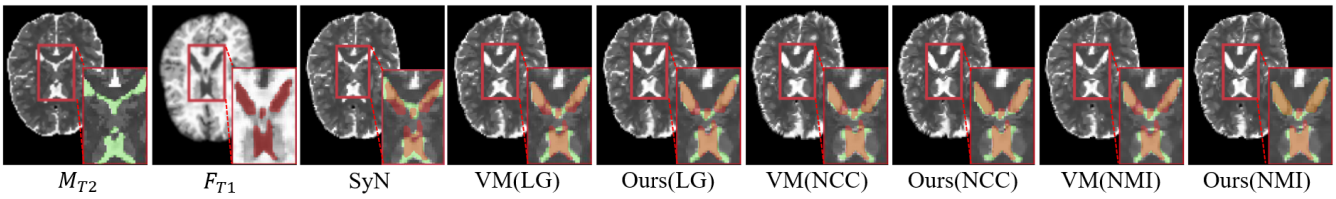

Fig. 2. Visualization of the registration results on the Brainweb. The green and red masks represent the anatomical structures of the warped moving image and the fixed image.

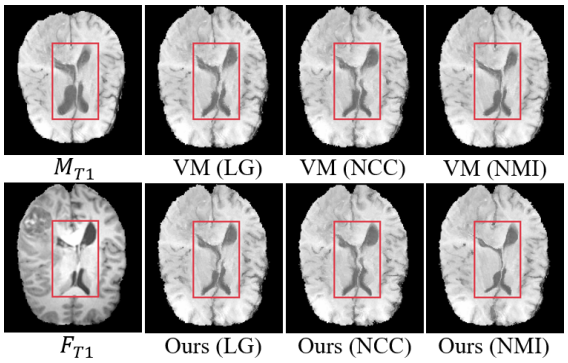

Fig. 3. Visualization of the registration results on the Brats.

Brainweb, which is trained under its default settings. PSNR and SSIM are leveraged to evaluate the quality of M_{T2-T1}. The quantitative results are provided in Table 3, which indicates that the proposed model using NMI loss obtains the best modal translation result among all the methods.

Table 3. Modal translation results on the Brainweb.

Method	PSNR	SSIM
CycleGAN [21]	24.030	0.9140
Ours (LG)	21.061	0.8492
Ours (NCC)	21.283	0.8723
Ours (NMI)	**25.277**	**0.9397**

Fig. 4 is the visualization results of the modal translation, and M_{T1} is the ground-truth. As can be seen in the red and yellow boxes, our model translates the appearance accurately with the preserving of structure. However, in the case of a small number of unpaired images, CycleGAN changes the anatomical structures. The proposed model using NMI loss produces the clearest anatomical structure contours among three loss functions.

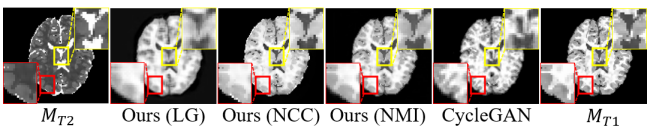

Fig. 4. Visualization of the modal translation results on the Brainweb.

3.4. Refinement of Deformation Field

Visualization of the refinement in the deformation field and its registration results are illustrated in Fig. 5. The three images in Fig. 5 (a) show the coarse, the fine and the final deformation field. Fig. 5 (b) visualizes the warped grids of the corresponding deformation fields. We use M_{T2} and F_{T1} for registration to illustrate the results of different deformation fields, which is shown in Fig. 5 (c). The area marked with the red box indicates that the final deformation field can achieve more accurate registration result after the refinement.

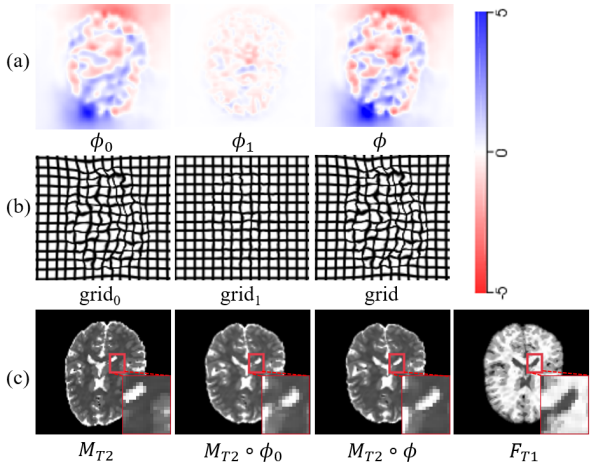

Fig. 5. Visualization of the deformation fields and the registration results. (a), (b) and (c) show the deformation fields, the warped grids and the registration results respectively.

4. CONCLUSION

In this paper, we propose an unsupervised hierarchical translation-based model for multi-modal medical image registration. The proposed model with three components can perform the modal translation and the registration of multi-modal images concurrently. Furthermore, by applying a staged training scheme in our method, the modal translation and registration results complement and benefit each other. The experiments on two multi-modal brain image datasets show that our method achieves significant improvement compared with the existing methods.

5. REFERENCES

[1] B. Avants, C. Epstein, M. Grossman, and J. Gee, "Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain," *Medical Image Analysis*, vol. 12, no. 1, pp. 26–41, 2008.

[2] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens, "Multimodality image registration by maximization of mutual information," *IEEE Transactions on Medical Imaging*, vol. 16, no. 2, pp. 187–198, 1997.

[3] G. Balakrishnan, A. Zhao, M. Sabuncu, J. Guttag, and A. Dalca, "An unsupervised learning model for deformable medical image registration," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9252–9260, 2018.

[4] S. Zhao, Y. Dong, E. Chang, Y. Xu, et al., "Recursive cascaded networks for unsupervised medical image registration," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10600–10610, 2019.

[5] X. Hu, M. Kang, W. Huang, M. Scott, R. Wiest, and M. Reyes, "Dual-stream pyramid registration network," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 382–390, 2019.

[6] M. Meng, L. Bi, M. Fulham, D. Feng, and J. Kim, "Enhancing medical image registration via appearance adjustment networks," *arXiv preprint arXiv:2103.05213*, 2021.

[7] B. de Vos, F. Berendsen, M. Viergever, H. Sokooti, M. Staring, and I. Išgum, "A deep learning framework for unsupervised affine and deformable image registration," *Medical Image Analysis*, vol. 52, pp. 128–143, 2019.

[8] C. Guo, *Multi-modal image registration with unsupervised deep learning*, Ph.D. thesis, Massachusetts Institute of Technology, 2019.

[9] J. Chen, J. Wei, and R. Li, "Targan: Target-aware generative adversarial networks for multi-modality medical image translation," *arXiv preprint arXiv:2105.08993*, 2021.

[10] B. Xin, Y. Hu, Y. Zheng, and H. Liao, "Multi-modality generative adversarial networks with tumor consistency loss for brain mr image synthesis," in *2020 IEEE 17th International Symposium on Biomedical Imaging (ISBI)*. IEEE, pp. 1803–1807, 2020.

[11] P. Huang, D. Li, Z. Jiao, D. Wei, G. Li, Q. Wang, H. Zhang, and D. Shen, "Coca-gan: common-feature-learning-based context-aware generative adversarial network for glioma grading," in *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, pp. 155–163, 2019.

[12] B. Yu, L. Zhou, L. Wang, Y. Shi, J. Fripp, and P. Bourgeat, "Ea-gans: edge-aware generative adversarial networks for cross-modality mr image synthesis," *IEEE Transactions on Medical Imaging*, vol. 38, no. 7, pp. 1750–1762, 2019.

[13] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical Image Computing and Computer-assisted Intervention*. Springer, pp. 234–241, 2015.

[14] M. Jaderberg, K. Simonyan, A. Zisserman, et al., "Spatial transformer networks," *Advances in Neural Information Processing Systems*, vol. 28, pp. 2017–2025, 2015.

[15] C. Cocosco, V. Kollokian, R. Kwan, G. Pike, and A. Evans, "Brainweb: Online interface to a 3d mri simulated brain database," in *NeuroImage*. Citeseer, 1997.

[16] B. Menze, A. Jakab, S. Bauer, J. Kalpathy-Cramer, K. Farahani, J. Kirby, Y. Burren, N. Porz, J. Slotboom, R. Wiest, et al., "The multimodal brain tumor image segmentation benchmark (brats)," *IEEE Transactions on Medical Imaging*, vol. 34, no. 10, pp. 1993–2024, 2014.

[17] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., "Tensorflow: A system for large-scale machine learning," in *12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16)*, pp. 265–283, 2016.

[18] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[19] B. Avants, N. Tustison, G. Song, et al., "Advanced normalization tools (ants)," *Insight j*, vol. 2, no. 365, pp. 1–35, 2009.

[20] X. Zhang, W. Jian, Y. Chen, and S. Yang, "Deform-gan: an unsupervised learning model for deformable registration," *arXiv preprint arXiv:2002.11430*, 2020.

[21] J. Zhu, T. Park, P. Isola, and A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2223–2232, 2017.