Multi-view Cross-Modality MR Image Translation for Vestibular Schwannoma and Cochlea Segmentation

Bogyeong Kang, Hyeonyeong Nam, Ji-Wung Han,
Keun-Soo Heo, and Tae-Eui Kam[*]

Department of Artificial Intelligence, Korea University, Seoul, Republic of Korea
{kangbk,kamte}@korea.ac.kr

Abstract. In this work, we propose a multi-view image translation framework, which can translate contrast-enhanced T_1 (ceT_1) MR imaging to high-resolution T_2 (hrT_2) MR imaging for unsupervised vestibular schwannoma and cochlea segmentation. We adopt two image translation models in parallel that use a pixel-level consistent constraint and a patch-level contrastive constraint, respectively. Thereby, we can augment pseudo-hrT_2 images reflecting different perspectives, which eventually lead to a high-performing segmentation model. Our experimental results on the CrossMoDA challenge show that the proposed method achieved enhanced performance on the vestibular schwannoma and cochlea segmentation..

Keywords: Multi-view image translation · Cross-modality · MRI segmentation · Unsupervised domain adaptation.

1 Introduction

Vestibular schwannoma (VS) is a benign tumor that occurs in the nerve membrane cells of the vestibular nerve [6, 11]. For diagnosis and treatment of VS, it is necessary to segment the VS and its surrounding organs, especially the cochleas [6, 11]. In general, VS is diagnosed through contrast-enhanced T_1 (ceT_1) MR imaging but there are concerns about side effects such as allergy to gadolinium-containing contrast agents [6, 11]. As an alternative, high-resolution T_2 (hrT_2) MR imaging, a non-contrast imaging technique, has shed light on VS segmentation [6, 11]. However, it is very time-consuming and expensive to manually annotate newly released data. For this reason, the lack of annotated data can be a big problem for applying deep learning techniques in the medical domain. This issue can be solved by applying unsupervised domain adaptation, which allows a model trained in one domain to be adapted in another unseen domain without supervision [1, 5, 8]. Recently, some studies [12, 4, 3] have been conducted based on cross-modality domain adaptation for VS and cochlea segmentation in unseen hrT_2 scans. Previous studies [12, 4, 3] achieved outstanding performance

[*] Corresponding author.

on VS and cochlea segmentation utilizing image translation models such as CycleGAN [14] or CUT [10]. Of note, CycleGAN employs pixel-level consistent constraints, while CUT adopts patch-level contrastive constraints. The former constraint can better reflect the intensity and the texture of VS through cycle-consistency loss, but the structure of VS and cochleas could be distorted. Besides, the latter constraint uses contrastive loss, having an advantage in preserving the structure of VS and cochleas, but could ignore the detailed characteristics such as intensity and texture. Based on these considerations, we believe that we can obtain diverse pseudo-hrT2 images, which can help to improve the segmentation model performance by using the two aforementioned constraint models together.

Therefore, we design a multi-view image translation framework to obtain the pseudo-hrT_2 images with different perspectives by adopting two image translation models in parallel, CycleGAN [14] and QS-Attn [7]. CycleGAN employs a pixel-level consistent constraint, and QS-Attn is an advanced patch-level contrastive constraint method that focuses on domain-relevant features [7]. To our best knowledge, QS-Attn [7] is first adopted for image translation from ceT_1 to hrT_2 images in this work. Based on our multi-view image translation framework, the following segmentation model can learn both structure and texture of VS and cochleas.

2 Related Work

Cross-modality unsupervised domain adaptation has drawn a lot of attention in the CrossMoDA challenge [6]. The goal of this challenge is to construct a VS and cochlea segmentation model on hrT_2 images with unpaired annotated ceT_1 and non-annotated hrT_2 scans. Recent studies [12, 4, 3] first translated the source ceT_1 images to the target hrT_2 images, and then trained their segmentation models with the translated hrT_2 (i.e., pseudo-hrT_2) images. More specifically, Shin et al. [12] translated the ceT_1 images to the hrT_2 images by adding an additional decoder to CycleGAN to preserve the structures of VS and cochleas. Dong et al. [4] conducted image translation using NiceGAN [2], which is based on CycleGAN [14], and Choi et al. [3] obtained pseudo-hrT_2 images using CUT [10]. Of note, they all obtained pseudo-hrT_2 images by taking only one constraint model. Besides, Choi et al. [3] performed post-processing to obtain the images with low intensity, similar to the VS in real hrT_2 scans.

3 Proposed Method

3.1 Overview

Fig. 1 shows an overview of our proposed framework, which consists of three parts; (1) multi-view image translation, (2) segmentation model training, and (3) self-training. Specifically, we first generate the pseudo-hrT_2 images with various characteristics through multi-view image translation. After that, we train the segmentation model using the multi-view pseudo-hrT_2 images and the labels

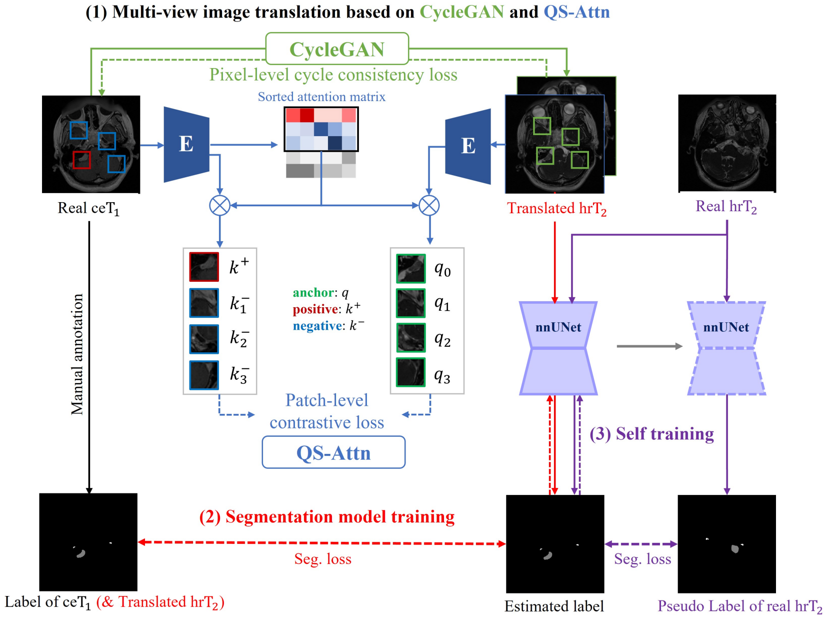

Fig. 1. The overview of our proposed framework.

of the ceT$_1$ images. In the self-training, the trained segmentation model first performs pseudo-labeling of real hrT$_2$ images, and then is further trained by including the pseudo-labeled real hrT$_2$ images in the next training phase.

3.2 Multi-view image translation

We first translate ceT$_1$ images into multi-view pseudo-hrT$_2$ images by adopting CycleGAN [14] and QS-Attn [7] in parallel.

CycleGAN. CycleGAN uses cycle-consistency loss to translate the source domain ceT$_1$ images into the target domain hrT$_2$ images. Cycle-consistency loss described in Eq. 1 encourages $F(G(x_s))$ to be equal to x_s and $G(F(x_t))$ to be equal to x_t in pixel-level when given the $G : X_s \rightarrow X_t$ and $F : X_t \rightarrow X_s$ generators [14].

$$L_{cycle} = \|F(G(x_s)) - x_s\| + \|G(F(x_t)) - x_t\| \tag{1}$$

QS-Attn. QS-Attn is an unpaired image translation model that is improved from CUT [10]. CUT preserves the structural information by constraining the patches from the same location on the source and the translated images to be close, compared to the different locations. CUT maximizes the mutual information between the source and translated images through the Eq. 2 [10],

$$L_{con} = -\log\left[\frac{\exp(q \cdot k^{+}/\tau)}{\exp(q \cdot k^{+}/\tau) + \sum_{i=1}^{N-1} \exp(q \cdot k^{-}/\tau)}\right] \tag{2}$$

where q is the anchor feature from the translated image and k^{+} is a single positive at the same location in the source image and k^{-} are $(N-1)$ negatives at the other locations, and τ is a temperature [7].

However, CUT [10] calculates the contrastive loss between the randomly selected patches, which could have less domain-relevant information. QS-Attn addresses this limitation by adopting the QS-Attn module, which can select domain-relevant patches. The QS-Attn module constructs the attention matrix A_g using the features in the source images and then obtains the entropy H_g by following Eq. 3 [7].

$$H_g(i) = -\sum_{j=1}^{HW} A_g(i,j) \log A_g(i,j) \tag{3}$$

Of note, the smaller entropy H_g means the more important feature. Thus, A_g is sorted in ascending order according to entropy H_g to select domain-relevant patches [7]. By calculating the contrastive loss using the selected domain-relevant patches, the structures of the source domain better preserve, and more realistic images are generated compared to CUT [10].

We empirically found that CycleGAN with pixel-level cycle-consistency loss allows the model to better reflect the intensity and the texture of the VS and cochleas in the target images, while QS-Attn takes advantage of preserving the structure of them more clearly via patch-level contrastive loss (refer to Section 5 for more details). By using them together, our multi-view image translation can augment pseudo-hrT$_2$ images from different perspectives, and it can help improve the performance of the following segmentation model.

3.3 Segmentation and Self-training

Motivated by the previous works [12, 4, 3], we also utilize nnUNet [9] and self-training procedure [13] to construct the segmentation model. nnUNet is a powerful segmentation framework that automatically performs pre-processing, training, and post-processing with heuristic rules [9]. Self-training is carried out to reduce the distribution gap between real hrT$_2$ and translated hrT$_2$ images and to improve the robustness of the segmentation model for unseen real hrT$_2$ scans. The segmentation and self-training procedure consists of four steps; (1) training the segmentation model using the translated hrT$_2$ scans with labels of the ceT$_1$ scans. (2) Generating pseudo labels of unlabeled real hrT$_2$ scans by using the trained segmentation model. (3) Retraining the segmentation model using both the translated hrT$_2$ scans with labels of the ceT$_1$ scans and the real hrT$_2$ scans with pseudo labels. 4) Repeating Steps 2-3 to achieve further performance improvement.

4 Experiments and Results

4.1 Dataset and preprocessing

We used the CrossMoDA dataset [1] [6] for training, validation. The CrossMoDA dataset consists of data from two different institutions: London and Tilburg. The London data consists of 105 ceT$_1$ scans and 105 hrT$_2$ scans. The ceT$_1$ scans were acquired with the in-plane resolution of 0.4×0.4mm, in-plane matrix of 512×512, and slice thickness of 1.0 to 1.5 mm with an MPRAGE sequence (TR=1900 ms, TE=2.97 ms, TI=1100 ms). Meanwhile, hrT$_2$ scans were acquired with the in-plane resolution of 0.5×0.5mm, in-plane matrix of 384×384 or 448×448, and slice thickness of 1.0 to 1.5 mm with a 3D CISS or FIESTA sequence (TR=9.4 ms, TE=4.23ms). For the Tilburg data set, ceT$_1$ scans and hrT$_2$ scans consist of 105 subjects each. The ceT$_1$ scans were acquired with the in-plane resolution of 0.8×0.8mm, in-plane matrix of 256×256, and slice thickness of 1.5 mm with a 3D-FFE sequence (TR=25 ms, TE=1.82 ms). The hrT$_2$ scans were acquired with the in-plane resolution of 0.4×0.4mm, in-plane matrix of 512×512, and slice thickness of 1.0 mm with a 3D-TSE sequence (TR=2700 ms, TE=160 ms, ETL=50) [6]. The training dataset of the CrossMoDA2022 Challenge [1] contains a total of 210 ceT$_1$ scans with annotation labels and 210 hrT$_2$ scans without annotation labels. In addition, they provide 64 scans of hrT$_2$ images for validation.

Since the voxel spaces vary across scans, all the images were resampled to [0.41, 0.41, 1.5] voxel sizes. For image translation, the 3D MRI images were sliced into a series of 2D images along the axial plane and the images were center-cropped and resized to 256 × 256. After performing image translation, the translated hrT$_2$ images were merged into 3D MR imaging and fed into the segmentation model.

4.2 Implementation details

We implement CycleGAN [14], QS-Attn [7], and nnUNet [9], following their default parameter settings. We also apply a global attention in QS-Attn [7], and ensemble selection in nnUNet [9] for the final prediction. All the implementations are powered by RTX 3090 24GB GPUs. The training of CycleGAN, QS-Attn, and nnUNet is performed with PyTorch 1.8.0, 1.7.1, and 1.10.2, respectively.

4.3 Results

Table 1 and Fig. 2 show the VS and cochlea segmentation results with different image translation methods. The proposed multi-view image translation framework with CycleGAN [14] and QS-Attn [7] shows better performance compared to other methods using each model alone. Moreover, we greatly improved the performance of the segmentation model with self-training. As a result, our proposed method obtained a great achievement with a mean dice score of 0.8504±0.0466 in the validation period.

[1] https://crossmoda-challenge.ml/

Table 1. Segmentation results with dice and ASSD scores (ST: self-training).

Translation model	Dice score			ASSD	
	VS	Cochlea	Mean	VS	Cochlea
CycleGAN (*w/o.* ST)	0.7798 (±0.1901)	0.8066 (±0.0323)	0.7932 (±0.0972)	0.8750 (±0.9222)	0.2422 (±0.1608)
QS-Attn (*w/o.* ST)	0.7779 (±0.1825)	0.8158 (±0.0287)	0.7968 (±0.0929)	0.6667 (±0.3891)	0.2365 (±0.1573)
Proposed (*w/o.* ST)	0.8043 (±0.1656)	0.8158 (±0.0289)	0.8101 (±0.0863)	0.5742 (±0.2461)	0.2387 (±0.1581)
Proposed (*w.* ST)	**0.8520 (±0.0889)**	**0.8488 (±0.0235)**	**0.8504 (±0.0466)**	**0.4748 (±0.2072)**	**0.1992 (±0.1524)**

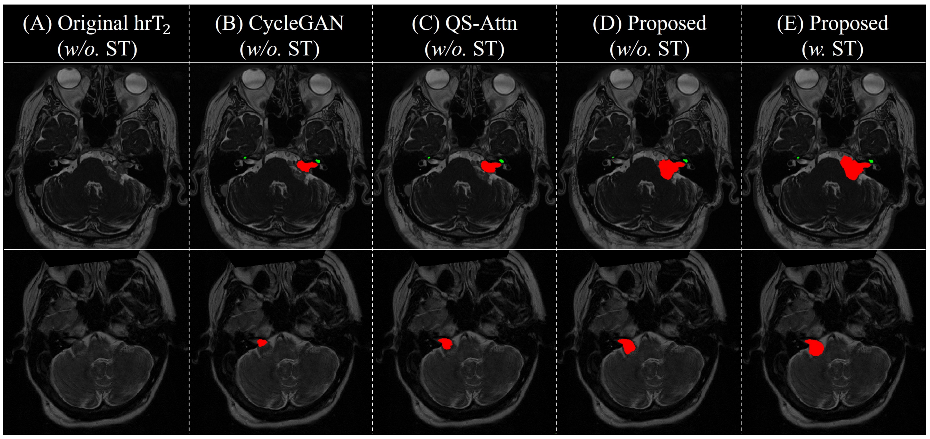

Fig. 2. Qualitative comparison of segmentation results for validation set. We visualize the segmentation results of VS (red) and cochlea (green) (ST: Self-training).

We conducted paired t-test among CycleGAN [14], QS-Attn [7], and our proposed method (*w/o.* self-training, ST) to compare the segmentation performance, and the results are plotted in Fig. 3. CycleGAN, QS-Attn, and our proposed method (*w/o.* ST) show statistical significance with $p < 0.05$ for the dice score of VS and mean values. In addition, our proposed method (*w/o.* ST) is statistically better with $p < 0.0001$ than CycleGAN on the dice score of cochleas. Through this statistical comparison, we proved that our proposed framework achieved better performance compared to other methods that use either of the two models alone.

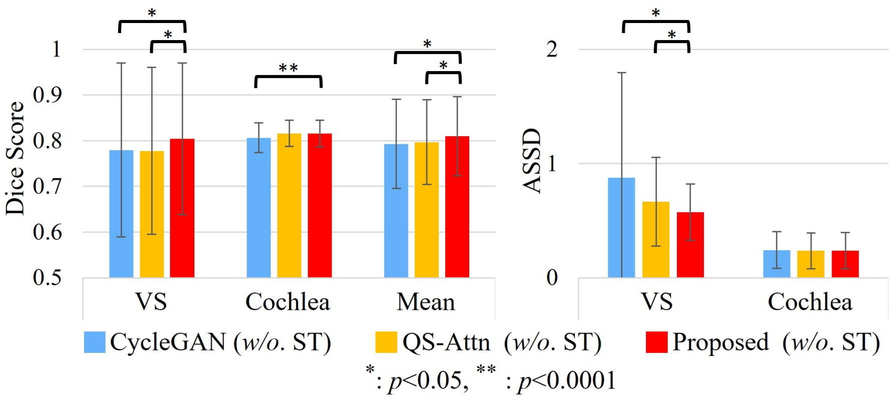

Fig. 3. Performance comparison of VS and cochlea segmentation models (ST: Self-training).

5 Discussion

Fig. 4 shows the results of the two separate image translation models utilized in the multi-view image translation framework. For comparison, we randomly picked two ceT$_1$ images (A&E), their corresponding translated hrT$_2$ images (B-C&F-G), and two unpaired real hrT$_2$ images (D&H). We can see that QS-Attn (C) well captured the structure of cochleas with less distortion or blurring compared to CycleGAN (B). Meanwhile, some images translated through QS-Attn (G) have too high intensities for VS, whereas those by CycleGAN (F) have similar intensity and textures to VS in the real hrT$_2$ image (H). As shown in Fig. 4, the two constraint models have different strengths. Therefore, in the proposed method, the segmentation model can learn both structures and textures of VS and cochleas through our multi-view image translation framework. It allows the segmentation model to consider various perspectives of VS and cochleas and helps improve the performance of the segmentation model.

6 Conclusion

In this work, we design a multi-view image translation framework for VS and cochlea segmentation. Specifically, we adopt CycleGAN and QS-Attn in parallel to translate the given ceT1 images to pseudo-hrT2 images reflecting various perspectives. Based on the pseudo-hrT2 images, the segmentation model can learn both structures and textures of VS and cochleas. Our proposed method obtained great achievement in the CrossMoDA challenge2022.

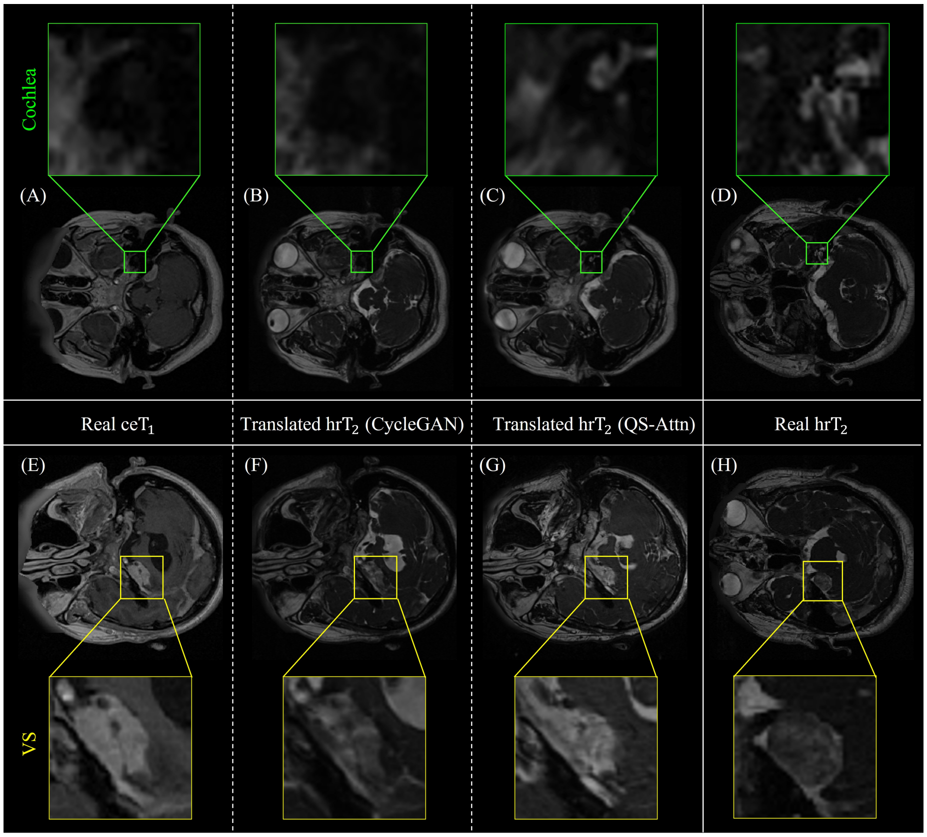

Fig. 4. Comparison results of image translation by CycleGAN and QS-Attn.

Acknowledgment. This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. 2019-0-00079, Artificial Intelligence Graduate School Program (Korea University)), and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. 2020R1C1C1013830, No. 2020R1A4A1018309).

References

1. Chen, C., Dou, Q., Chen, H., Qin, J., Heng, P.A.: Synergistic image and feature adaptation: Towards cross-modality domain adaptation for medical image segmentation. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 865–872 (2019)
2. Chen, R., Huang, W., Huang, B., Sun, F., Fang, B.: Reusing discriminators for encoding: Towards unsupervised image-to-image translation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8168–8177 (2020)

3. Choi, J.W.: Using out-of-the-box frameworks for unpaired image translation and image segmentation for the crossmoda challenge. arXiv e-prints pp. arXiv–2110 (2021)
4. Dong, H., Yu, F., Zhao, J., Dong, B., Zhang, L.: Unsupervised domain adaptation in semantic segmentation based on pixel alignment and self-training. arXiv preprint arXiv:2109.14219 (2021)
5. Dorent, R., Joutard, S., Shapey, J., Bisdas, S., Kitchen, N., Bradford, R., Saeed, S., Modat, M., Ourselin, S., Vercauteren, T.: Scribble-based domain adaptation via co-segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 479–489. Springer (2020)
6. Dorent, R., Kujawa, A., Ivory, M., Bakas, S., Rieke, N., Joutard, S., Glocker, B., Cardoso, J., Modat, M., Batmanghelich, K., et al.: Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwnannoma and cochlea segmentation. arXiv preprint arXiv:2201.02831 (2022)
7. Hu, X., Zhou, X., Huang, Q., Shi, Z., Sun, L., Li, Q.: Qs-attn: Query-selected attention for contrastive learning in i2i translation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18291–18300 (2022)
8. Huo, Y., Xu, Z., Moon, H., Bao, S., Assad, A., Moyo, T.K., Savona, M.R., Abramson, R.G., Landman, B.A.: Synseg-net: Synthetic segmentation without target modality ground truth. IEEE transactions on medical imaging **38**(4), 1016–1025 (2018)
9. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
10. Park, T., Efros, A.A., Zhang, R., Zhu, J.Y.: Contrastive learning for unpaired image-to-image translation. In: European conference on computer vision. pp. 319–345. Springer (2020)
11. Shapey, J., Kujawa, A., Dorent, R., Wang, G., Dimitriadis, A., Grishchuk, D., Paddick, I., Kitchen, N., Bradford, R., Saeed, S.R., et al.: Segmentation of vestibular schwannoma from mri, an open annotated dataset and baseline algorithm. Scientific Data **8**(1), 1–6 (2021)
12. Shin, H., Kim, H., Kim, S., Jun, Y., Eo, T., Hwang, D.: Cosmos: Cross-modality unsupervised domain adaptation for 3d medical image segmentation based on target-aware domain translation and iterative self-training. arXiv preprint arXiv:2203.16557 (2022)
13. Xie, Q., Luong, M.T., Hovy, E., Le, Q.V.: Self-training with noisy student improves imagenet classification. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10687–10698 (2020)
14. Zhu, J.Y., Park, T., Isola, P., Efros, A.A.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)