scientific reports

OPEN

Target-aware cross-modality unsupervised domain adaptation for vestibular schwannoma and cochlea segmentation

Bogyeong Kang[1], Hyeonyeong Nam[1], Myeongkyun Kang[2], Keun-Soo Heo[1], Minjoo Lim[1], Ji-Hye Oh[1] & Tae-Eui Kam[1]✉

There is growing interest in research on segmentation for the vestibular schwannoma (VS) and cochlea using high-resolution T2 (hrT2) imaging over contrast-enhanced T1 (ceT1) imaging due to the contrast agent side effects. However, the hrT2 imaging remains a problem of insufficient annotated data, which is fatal for building more robust segmentation models. To address the issue, recent studies have adopted unsupervised domain adaptation approaches that translate ceT1 images to hrT2 images. However, previous studies did not consider the size and visual characteristics of the target objects, such as VS and cochlea, during image translation. Specifically, those works simply performed normalization on the entire image without considering its significant impact on the quality of the translated images. These approaches tend to erase the small target objects, making it difficult to preserve the structure of these objects when generating pseudo-target images. Furthermore, they may also struggle to accurately reflect the unique style of the target objects within the images. Therefore, we propose a target-aware unsupervised domain adaptation framework, designed for translating target objects, each tailored to their unique visual characteristics and size using target-aware normalization. We demonstrate the superiority of the proposed framework on a publicly available challenge dataset. Codes are available at https://github.com/Bokyeong-Kang/TANQ.

A vestibular schwannoma (VS) is a benign tumor that develops in the nerve sheath cells of the vestibular nerve[1]. As the VS grows, it compresses the adjacent cranial nerves and blood vessels, leading to the deterioration of hearing and vestibular functions[2]. Moreover, it can cause hydrocephalus by obstructing the circulation of the cerebrospinal fluid[3] and develop severe adhesions with adjacent organs, such as the cochlea. Hence, it is essential to diagnose and treat a VS promptly to prevent adhesions and potential damage to associated organs[4]. In the diagnosis of a VS, techniques that accurately segment VS and surrounding organs[5], especially the cochleas, play a key role in determining the severity of VS[6].

Magnetic resonance imaging has been widely used for VS segmentation because it can clearly determine the degree of adhesion with surrounding organs, such as the cochleas[4]. In particular, contrast-enhanced T_1 (ceT_1) magnetic resonance imaging has traditionally been used for its capability to precisely visualize neural structures using a gadolinium-containing contrast agent[7]. Recently, owing to the potential risks associated with gadolinium-based contrast agents[8], high-resolution T_2 (hrT_2) imaging has gained attention as an alternative to ceT_1 imaging with far lower associated costs and risks[8].

However, because the hrT_2 imaging modality has been introduced for VS segmentation recently, it suffers from a relative lack of annotation labels for VS and cochleas compared with the ceT_1 imaging modality[1]. Notably, acquiring the annotation labels for hrT_2 images incurs additional costs because of the time-consuming and labor-intensive nature of manual annotation conducted by radiologists and physicians[9]. This issue significantly limits the use of deep learning (DL)-based models for automatic VS segmentation with hrT_2 images because the small amount of annotated data used for model training significantly degrades model performance and reliability[10].

To address the label scarcity in hrT_2 imaging, numerous studies[11–14] have been proposed; these employ unsupervised domain adaptation (UDA) by translating imaging modalities from a label-rich source domain (*i.e.*, ceT_1) to a label-poor target domain (*i.e.*, hrT_2). This approach allows the annotation labels from the

[1]Department of Artificial Intelligence, Korea University, Seoul, South Korea. [2]Department of Robotics and Mechatronics Engineering, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, South Korea. ✉email: kamte@korea.ac.kr

source domain to be directly utilized for the target domain to train a segmentation model for VS and cochleas based on target $\mathrm{hrT_2}$ images. Specifically, in this approach, pseudo-$\mathrm{hrT_2}$ images from real $\mathrm{ceT_1}$ images are first generated by translating the visual characteristics (or style) of $\mathrm{ceT_1}$ images (*i.e.,* source domain) to those of $\mathrm{hrT_2}$ images (*i.e.,* target domain) while preserving the anatomical structures of the source $\mathrm{ceT_1}$ images. Subsequently, pairs of the generated pseudo-$\mathrm{hrT_2}$ images and annotation labels on $\mathrm{ceT_1}$ images are used to train a DL-based segmentation model for VS and cochleas based on $\mathrm{hrT_2}$ images. Finally, the VS and cochleas in the real $\mathrm{hrT_2}$ images are segmented with the trained segmentation model.

Earlier studies[11,12] utilized basic image translation models *e.g.,* NiceGAN[15] and CUT[16]. However, these methods focused on generating pseudo-images that simply resembled the style of the target domain, rather than performing translation with a focus on the target object *e.g.,* the VS and cochlea. Shin *et al.*[13] and Dong *et al.*[14] have enhanced these methods by integrating a segmentation decoder into the generator. This refined the generator training through the segmentation output, thereby effectively preserving the structural integrity of the VS and cochlea. In our previous study[17], we further developed the UDA approach to preserve the VS and cochlea. Our method employed the query-selected attention (QS-Attn) model[18], which was designed to focus on the informative regions of the source images during translation. Our method successfully preserved the structural fidelity of the VS and cochlea during image translation with contrastive learning. Our approach is effective in focusing on small and critical target objects during translation. Additionally, we adopted a multi-view approach to generate diverse images by leveraging two constraint models with distinct strengths to enhance the performance of the segmentation model. As a result, our method achieved second place in the CrossMoDA challenge 2022 with the best and second-best segmentation results for VS and cochleas, respectively.

Despite their success, existing methods[11–14,17] do not consider the size and visual characteristics of the target object during image translation, both of which are important for successful image translation from $\mathrm{ceT_1}$ to $\mathrm{hrT_2}$. Most existing methods[11–13], including our primary work[17], have adopted instance normalization (IN)[19], which performs normalization within the entire image. While IN can be effective for image synthesis tasks, it has limitations for our specific task due to the small size and distinct visual characteristics of the VS and cochlea in $\mathrm{ceT_1}$ and $\mathrm{hrT_2}$ images. In contrast to the target objects in other medical imaging datasets such as Multi-Modality Whole Heart Segmentation (MMWHS)[20], the VS and cochlea in $\mathrm{ceT_1}$ and $\mathrm{hrT_2}$ are extremely small, accounting for only 0.028% and 0.002% of the total volume, respectively[10]. Furthermore, there are significant differences in the image characteristics of $\mathrm{ceT_1}$ and $\mathrm{hrT_2}$ for these objects. As show in Fig. 1 (a), in $\mathrm{ceT_1}$ images, the VS voxels appear with higher intensity compared to other tissues due to the contrast agent. Conversely, as show in Fig. 1 (b), in $\mathrm{hrT_2}$ images, the VS is less distinguishable from surrounding tissues since the voxels of the VS have low intensity similar to surrounding tissues[1,17,21]. Additionally, the cochlea, which is even smaller, appears at a very low intensity in $\mathrm{ceT_1}$ images (Fig. 1 (a)) compared with the $\mathrm{hrT_2}$ images (Fig. 1 (b))[1,17,21]. For these reasons, IN contains two major limitations for our task of translating $\mathrm{ceT_1}$ to $\mathrm{hrT_2}$ images. Firstly, it tends to "wash away"[22] the structural information of extremely small-sized target objects such as the VS and cochlea, with the cochlea being particularly affected due to its low intensity. This "wash away" effect is a significant limitation due to the failure in generating pseudo-$\mathrm{hrT_2}$ images that preserve the detailed structures of the VS and cochlea. Secondly, it is challenging to reflect the low-intensity style of VS that appears in real $\mathrm{hrT_2}$ images. Considering the significant style differences, such as intensity, between the VS in source and target images, tailored adjustments are necessary to reflect the low-intensity style of the VS in pseudo-$\mathrm{hrT_2}$ images. Therefore, only applying IN without these specific adjustments can lead to a continuous presence of the high-intensity style of VS, originally observed in the source images, which is not desirable for the target images.

To alleviate these issues, we propose a novel target-aware, UDA framework that integrates a Target-Aware Normalization-based QS-Attn(TANQ). The TANQ is designed to facilitate image translation of target objects, each tailored to their distinct visual characteristics in $\mathrm{ceT_1}$ and $\mathrm{hrT_2}$ images. Specifically, it is designed to prevent the "wash away" effect that erases details in the small and low-intensity cochlea. We separate the cochlea region

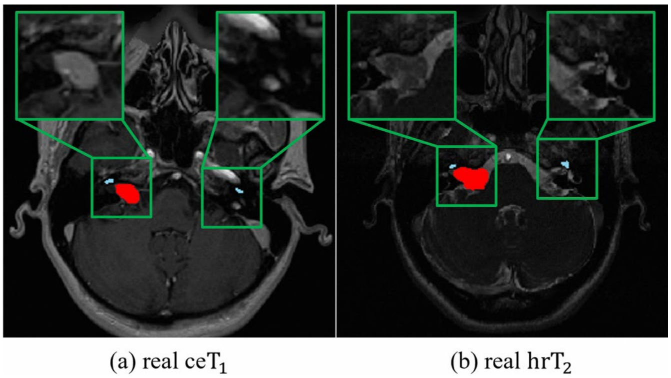

Fig. 1. Difference in characteristics of target objects (VS and cochlea) between $\mathrm{ceT_1}$ and $\mathrm{hrT_2}$ images. (**a**) represents a real $\mathrm{ceT_1}$ image, and (**b**) is a real $\mathrm{hrT_2}$ image, with the VS highlighted in red and the cochlea in sky-blue by overlapping the segmentation map. In (**a**) real $\mathrm{ceT_1}$, the segmentation map was overlaid using the real annotation labels, while in (**b**) real $\mathrm{hrT_2}$ image, since real annotation labels were not available, we used pseudo-labels for the overlap.

from others and perform individual normalization of its features within the region. This approach ensures the preservation of the cochlea's fine structural details. Additionally, to address the issue of the high intensity of the VS within pseudo-hrT_2 images, we inject the style information from surrounding tissues into the VS during its normalization, reflecting its low-intensity in pseudo-hrT_2 images. This adjustment aims to match the VS's style to the lower intensity observed in real target images. Based on target-aware normalization techniques, TANQ can enhance its ability to: 1) preserve the structural information of each target object, even for very small and low-intensity cochleas; and 2) convert the target objects to possess specific visual characteristics such as textures and intensities typically observed in real hrT_2 images. The TANQ model enables the generation of realistic pseudo-hrT_2 images with accurately preserved structures of the target objects while ensuring that the VS exhibits low-intensity characteristics, resulting in a direct enhancement of the segmentation model performance. TANQ is then integrated with our primary framework comprising multi-view image translation and self-training-based segmentation[17] to further improve the performance of VS/cochlea segmentation based on unannotated hrT_2 images.

The key contributions of our study are outlined as follows:

- We propose a target-aware unsupervised domain adaptation framework, aimed at translating target objects, with each tailored to their specific visual characteristics and size through target-aware normalization.
- Based on our in-depth analysis of visual characteristics between source ceT_1 and target hrT_2 modalities, we prevented structural loss of target objects and generated higher quality pseudo-hrT_2 images, thereby enhancing the performance of the VS and cochlea segmentation models.
- We demonstrate the superiority of the proposed method compared with other existing methods and the effectiveness of our framework through quantitative and qualitative analysis using a publicly available challenge dataset.

Related work

Unsupervised domain adaptation addresses the domain shift between source and target domains, enabling it to perform a task better in the target domain without the need for annotation labels. Several studies have adopted feature alignment[23,24], which mitigates domain discrepancies by mapping the source and target images to a shared latent space and learning domain-invariant representations through adversarial learning. Chen *et al.*[25] utilized a low-level feature refinement module and prediction map alignment to mitigate the domain gap at the feature level. In contrast, image-level alignment has been also adopted to translate the source domain into the target domain using a generative model and reducing the differences between the source and target domains in pixel-level space[26,27]. Kang *et al.*[28] applied mutual information and texture co-occurrence loss to convert the source domain into the target domain, successfully performing domain adaptation for segmentation tasks. Moreover, several research studies have attempted to combine image- and feature-level alignments by generating pseudo-target images first and then applying feature-level techniques to minimize the domain gap between pseudo- and real target images[29,30]. Hu *et al.*[31] proposed a domain-specific convolution module to extract domain-invariant features and employed a high-frequency reconstruction module to generate target images, thereby achieving domain adaptation.

In the medical domain, domain adaptation has been conducted to enhance the performance of downstream tasks such as segmentation, with a particular emphasis on preserving the structure of small target objects to achieve better results. Chen *et al.*[29] leveraged synergistic learning to apply both image and feature alignment, enabling bidirectional cross-modality domain adaptation between MRI and CT images. Jiang *et al.*[32] introduced a structure discriminator to model the co-dependency between images and their corresponding segmentations using joint probability, which helped focus on generating accurate representations of target objects. However, the target objects in these studies, such as cardiac structures and abdominal organs, are significantly larger than our target objects such as VS and cochlea. Given these distinct characteristics, there is growing recognition of the necessity for more intricate and refined processing through image-level alignment, which is now being preferred over feature-level[23,24] or combination approaches[29,30] in ongoing research.

In the CrossMoDA challenge, various image-level alignment approaches were utilized to translate ceT_1 images to hrT_2 images[11–13]. Specifically, Dong *et al.*[11] and Choi *et al.*[12] generated target pseudo-hrT_2 images from source ceT_1 images by utilizing NiceGAN[15] and CycleGAN[33], respectively. Shin *et al.*[13] and PAST[14] appended a segmentation decoder to a decoder within a CycleGAN-based generator and utilized the segmentation results in the image translation process to train the generator, thus improving the structural preservation of the VS and cochlea during image translation. Furthermore, several studies[17,21,34] have enhanced the richness of pseudo-target images using a range of image translation methods for augmentation. Fgh365[21] utilized cross-site and cross-modality image translation approaches using CycleGAN[33] to generate diverse pseudo-hrT_2 images with a rule-based offline augmentation technique for domain gap mitigation. Our prior multi-view approach[17] also attempted to mitigate the domain gap between ceT_1 and hrT_2 images by generating pseudo-hrT_2 images using two parallel constraint models. TBA[34] employed CycleGAN and tumor-blending augmentation with SinGAN[35] to enhance the appearance diversity of the regions of interest (e.g., the VS and cochlea) and achieve effective domain adaptation. However, despite these efforts, the structural information of the VS and cochlea was not perfectly preserved, and there still remains room for improvement in reflecting the style of VS and cochlea in real hrT_2 images.

Proposed method

As illustrated in Fig. 2, our proposed framework consists of three main parts: 1) TANQ-based image translation from ceT_1 to hrT_2 images, 2) multi-view pseudo-hrT_2 representation obtained using additional CycleGAN-

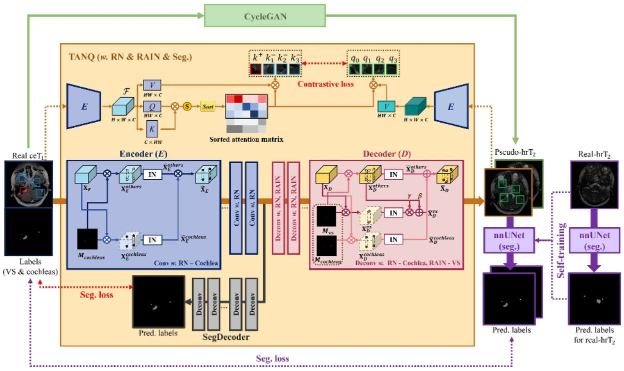

Fig. 2. Our proposed framework consists of three main parts: 1) TANQ-based image translation from $\mathrm{ceT_1}$ to $\mathrm{hrT_2}$ images, 2) Multi-view pseudo-$\mathrm{hrT_2}$ representation via CycleGAN, 3) Construction of a VS/cochlea segmentation model using multi-view pseudo-$\mathrm{hrT_2}$ images and self-training with real-$\mathrm{hrT_2}$ images. Specifically, TANQ divides the features based on the $\mathrm{ceT_1}$ labels in both the encoder and decoder, applying target-aware normalization. Furthermore, it includes an additional decoder called SegDecoder. The Encoder E extracts features from both the real $\mathrm{ceT_1}$ images and pseudo-$\mathrm{hrT_2}$ images and then calculates the contrastive loss between selected features using a sorted attention matrix.

based image translation in parallel, and 3) a VS/cochlea segmentation model construction based on multi-view pseudo-$\mathrm{hrT_2}$ images and self-training with real-$\mathrm{hrT_2}$ images.

TANQ-based image translation

Target-aware encoder–decoder construction

In the proposed TANQ, we constructed a ResNet-based encoder–decoder architecture, followed by the original QS-Attn[18]. Here, instead of IN[19] used in the original QS-Attn, which performs normalization using the mean and variance of the feature maps of each convolution layer in the architecture, we design a target-aware normalization approach in TANQ, which normalizes the feature maps of each convolution layer by focusing on the target objects (i.e., regions of VS and cochleas) in different processes according to their inherent visual characteristics in $\mathrm{ceT_1}$ and $\mathrm{hrT_2}$ images, as mentioned in Section 1. Specifically, as shown in Fig. 2, we first perform region normalization (RN)[36] for cochleas to prevent the loss of their anatomical structures, that is, the cochlear regions are normalized independently of other tissues as they exhibit relatively low-intensity textures and blurry boundaries in $\mathrm{ceT_1}$ images. By contrast, for the VS, we use region-aware adaptive instance normalization (RAIN)[37] in each decoder layer to convert a relatively high-intensity style of the VS in $\mathrm{ceT_1}$ images to the VS in the $\mathrm{hrT_2}$ images, which exhibits an intensity similar to that of other tissues.

Each feature map passing through the convolution in each encoder layer of TANQ can be described as: $\mathbf{X}_E \in \mathbb{R}^{H\times W\times C}$, where H, W, and C denote the height, width, and number of channels of each feature map, respectively. Based on the annotation mask of cochleas $\mathbf{M}^{cochleas}$ in $\mathrm{ceT_1}$ image, $\mathbf{X}_E$ can be divided into two categorical regions $r = \{cochleas, others\}$ as follows:

$$\mathbf{X}_E = \mathbf{X}_E^{cochleas} \cup \mathbf{X}_E^{others}, \tag{1}$$

where $\mathbf{X}_E^r$ represents the masked feature map of each categorical region of r. By adopting RN[36], each masked feature map $\mathbf{X}_E^r$ is separately normalized by region-wise IN, as shown in equation (2).

$$\widetilde{\mathbf{X}}_E^r = \frac{1}{\sigma^r}(\mathbf{X}_E^r - \mu^r), \tag{2}$$

where μ^r and σ^r denote the mean and standard deviation of each masked feature map $\mathbf{X}_E^r$, respectively. The normalized feature maps from all the regions are merged into $\widetilde{\mathbf{X}}_E$ as equation (3) and then passed as input to the next encoder layer.

$$\widetilde{\mathbf{X}}_E = \widetilde{\mathbf{X}}_E^{cochleas} \cup \widetilde{\mathbf{X}}_E^{others} \tag{3}$$

According to RN, which can retain the distinct characteristics of the cochleas in $\mathrm{ceT_1}$ images, the anatomical structures of the cochleas can be precisely preserved during the encoding process, and the information is passed to the subsequent decoding process as well as the contrastive learning process of QS-Attn[18] to generate more realistic pseudo-$\mathrm{hrT_2}$ images.

In the decoding process, each feature map X_D, which is passed through deconvolution in each decoder layer of TANQ, is divided into three different categorical regions $s = \{vs, cochleas, others\}$ based on the annotation masks of $\mathrm{ceT_1}$, that is, $\mathbf{M}^{vs}$ and $\mathbf{M}^{cochleas}$ for VS and cochleas, respectively:

$$\mathbf{X}_D = \mathbf{X}_D^{vs} \cup \mathbf{X}_D^{cochleas} \cup \mathbf{X}_D^{others}, \tag{4}$$

where $\mathbf{X}_D^s$ represents the masked feature map of each categorical region of *s*. Each region-wise feature map is also separately normalized based on region-wise IN, as shown in equation (5).

$$\widetilde{\mathbf{X}}_D^s = \frac{1}{\sigma^s}(\mathbf{X}_D^s - \mu^s), \tag{5}$$

where μ^s and σ^s denote the mean and standard deviation of each masked feature map $\mathbf{X}_D^s$, respectively. For VS, we further adopt RAIN[37] to render VS with an intensity similar to that of other tissues in pseudo-hrT_2 images as follows:

$$\overline{\mathbf{X}}_D^{vs} = \widetilde{\mathbf{X}}_D^{vs} \times \gamma^{others} + \beta^{others}, \tag{6}$$

where β^{others} and γ^{others} denote the mean and standard deviation of $\mathbf{X}_D^{others}$, respectively.

The normalized feature maps from all the regions are merged into $\overline{\mathbf{X}}_D$ as equation (7) and then passed as input to the next decoder layer, as shown in Fig. 2.

$$\overline{\mathbf{X}}_D = \overline{\mathbf{X}}_D^{vs} \cup \widetilde{\mathbf{X}}_D^{cochleas} \cup \widetilde{\mathbf{X}}_D^{others} \tag{7}$$

Furthermore, in TANQ, we introduce an additional decoding process to enhance its ability to preserve the structural information of the target objects in the encoding process, that is, the SegDecoder in Fig. 2. SegDecoder is responsible for generating segmentation outputs for the target objects, (VS and cochlea) using encoded feature maps. By adding SegDecoder, the attention of the encoder is directed toward the target objects, facilitating a more efficient extraction of feature maps while simultaneously preserving their structures. SegDecoder minimizes cross-entropy loss, which is calculated by comparing the output of the SegDecoder with the ground-truth annotation labels of ceT_1 image, as presented in the following equation[38]:

$$L_{seg} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_{i,c} \cdot \log\left(p_{i,c}\right), \tag{8}$$

where *N* is the total number of pixels, *C* is the total number of classes, $y_{i,c}$ is the ground-truth label for the i^{th} pixel and c^{th} class, which is one-hot encoded, and $p_{i,c}$ represents the predicted probability of the i^{th} pixel belonging to the c^{th} class.

Contrastive and adversarial learning in TANQ

In QS-Attn[18], contrastive learning forces a specific patch from the translated image to be close to the patch located at the same position in the source image and farther away from patches at different positions in the source image. This process enables the generator to preserve its structure while being insensitive to style differences. Unlike CUT[16], which selects patches randomly, QS-Attn identifies the informative patches used for contrastive learning according to the entropy of the features from the source images. To calculate the entropy of the features, the feature $\mathscr{F} \in \mathbb{R}^{H\times W\times C}$ is extracted from the ceT_1 images using encoder *E* in QS-Attn. The feature $\mathscr{F}$ is reshaped into a 2D matrix $Q \in \mathbb{R}^{HW\times C}$, multiplied by its transpose matrix $K \in \mathbb{R}^{C\times HW}$, and converted into the final attention matrix A_g by applying the softmax function. The entropy H_g of each row in A_g is then computed using equation (9), which allows us to measure the degree of similarity to other features[18].

$$H_g(i) = -\sum_{j=1}^{HW} A_g(i,j)\log A_g(i,j) \tag{9}$$

A low entropy value H_g indicates that only a few features are similar to the i^{th} query. This assumes a significant role in the calculation of contrastive loss as it distinctly sets them apart from other features. These features are prioritized by sorting the rows of A_g based on entropy, and the *N* smallest rows are used for the contrastive loss[18].

This contrastive learning of QS-Attn is further enhanced by combining it with the target-aware encoder–decoder architecture introduced in Section3.1 within a unified framework. Specifically, in TANQ, based on the target-aware encoder–decoder architecture designed to preserve the structures of target objects more precisely, patches around the target objects are more frequently selected as informative patches compared to the original QS-Attn (refer to Fig. 5 for more details). Consequently, contrastive learning applied to the selected patches around the target objects can attempt to restore their information in the source ceT_1 to pseudo-hrT_2 images during image translation in the proposed TANQ method, as shown in Fig. 2.

The full objective loss of TANQ is as follows:

$$L_G = L_{adv} + L_{con}^X + L_{con}^Y + L_{seg}, \tag{10}$$

where L_{adv} is the adversarial loss[39] used to generate pseudo-images similar to real hrT_2 images[16]. L^X_{con} refers to the contrastive loss between the source ceT_1 image and the translated hrT_2 image, and L^Y_{con} denotes the identity loss, which is calculated as the contrastive loss between the real target hrT_2 image and the translated hrT_2 image[16].

Multi-view representation

We employed a multi-view image translation approach based on our prior study[17]. Specifically, we improved the segmentation model performance by augmenting the pseudo-hrT_2 images using TANQ and CycleGAN[33] in parallel, as shown in Fig. 2. CycleGAN is well-suited for generating pseudo-hrT_2 images that reliably reflect the characteristics observed in real hrT_2 images owing to its use of a pixel-level reconstruction constraint[17]. Specifically, CycleGAN preserves the structural information of the ceT_1 image by enforcing cycle-consistency loss[33] as follows:

$$L_{cycle} = \|F(G(I_s)) - I_s\| + \|G(F(I_t)) - I_t\|\,, \tag{11}$$

where *G* and *F* indicate the generators from the target to the source images, and vice versa[33]. When a source image I_s, for example, ceT_1 image, is fed into the generator *G* and subsequently reconstructed by the generator *F*, the structure of the ceT_1 image is preserved by minimizing the discrepancy between $F(G(I_s))$ and I_s. Similarly, the structural information of the target image I_t, for example, the hrT_2 image, is preserved by minimizing the discrepancy between the reconstructed image $G(F(I_t))$ and original I_t. Consequently, CycleGAN also facilitates image translation from ceT_1 to the pseudo-hrT_2 images with the overall loss function of CycleGAN as follows:

$$L_{CycleGAN} = L^X_{adv} + L^Y_{adv} + L_{cycle}, \tag{12}$$

where L^X_{adv} represents the adversarial loss[39] used during the generation of hrT_2 images from ceT_1 images, while L^Y_{adv} represents the adversarial loss employed during the generation of ceT_1 images from hrT_2 images in the reverse process[33].

By employing our multi-view image translation approach, which combines the pseudo-hrT_2 images generated using TANQ and CycleGAN, we enhance the variability and richness of the pseudo-hrT_2 images. This approach enables a subsequent segmentation model to effectively learn the diverse characteristics present within hrT_2 images, thereby facilitating effective learning.

Segmentation model construction and self-training

We first construct a nnUNet[40]-based model for VS/cochlea segmentation with pseudo-hrT_2 images. The choice of nnUNet is based on its established performance in numerous medical image segmentation challenges as it automatically configures a UNet-based segmentation pipeline based on the analysis of the provided training cases[40]. We directly employed all multi-view pseudo-hrT_2 images, along with their associated annotation labels from the source ceT_1 images, to train the VS/cochlea segmentation model.

The segmentation model trained on the pseudo-hrT_2 images enhances its generalization ability for the unseen real hrT_2 images based on self-training[41]. The self-training process consists of the following four steps. 1) We train the segmentation model using multi-view pseudo-hrT_2 scans along with the ground-truth labels from the ceT_1 scans. 2) The segmentation model trained in step 1 is then utilized to generate pseudo-labels for the unlabeled real hrT_2 scans. 3) The segmentation model is retrained using both the pseudo-hrT_2 scans with ground-truth ceT_1 labels and the real hrT_2 scans with pseudo-labels. 4) To further enhance the performance of the segmentation model, we repeat steps 2 and 3 four times. This self-training process aims to refine the performance of the segmentation model by leveraging a combination of pseudo- and real hrT_2 images, that is, as the segmentation model is retrained, the pseudo-annotation labels for the real hrT_2 images are improved, and the results can be used again to further improve the model performance[41]. Although we employed a basic self-training approach, it is widely recognized that even simple self-training methods can enhance the performance and generalization of segmentation models[42].

Experimental settings

Dataset and preprocessing

We used the CrossMoDA dataset[1,4] to evaluate the effectiveness of our framework for VS/cochlea segmentation in unseen real hrT_2 scans. The dataset comprised 210 annotated ceT_1 and 210 unannotated hrT_2 scans to be used for training. We evaluated our framework using the 64 scans provided as the CrossMoDA validation dataset, and the evaluation was conducted using the official CrossMoDA 2022 website. The segmentation results of the dataset comprised the VS, cochlea, and background classes. The CrossMoDA dataset consisted of two distinct datasets: London SC-GK data and Tilburg SC-GK data[1], both of which provide an equal number of ceT_1 and hrT_2 scans. We resampled all ceT_1 scans to 0.41×0.41×1.5 voxel sizes to ensure consistency. The 3D scans were converted into a series of 2D images along the axial plane. The images were then center-cropped and resized to 256×256 pixels for image translation. Following image translation, the translated hrT_2 images were reconstructed into 3D volumes and fed into the proposed segmentation model.

Competing methods and ablation study

The competing methods[13,14,17,21,34,43] were derived from the top-ranked models of the CrossMoDA 2021, 2022 challenge, including our previous work, which secured the second position in the challenge. Furthermore, we conducted an ablation study to examine the effectiveness of the individual components of image translation within our framework. We trained segmentation models for the VS and cochlea using pseudo-hrT_2 images generated using each image translation method and evaluated their performance without conducting self-training. We also demonstrate the effect of target-aware normalization through the image translation results of target objects and visualization of informative patch selection in Section 5.3. The results were evaluated on the validation set using two evaluation metrics: the Dice score[44] and average symmetric surface distance (ASSD)[45]. Furthermore, we conducted a paired *t*-test to assess the statistical significance of our results, particularly focusing on the Dice score and ASSD metrics.

Implementation details

For the cross-modality image translation from ceT_1 to hrT_2 images, we employed CycleGAN[33] and TANQ, where both architectures consisted of a ResNet-based generator[46] and a PatchGAN-based discriminator[47]. We trained CycleGAN and TANQ with a batch size of 4 for 200 epochs and 400 epochs, respectively, using the Adam[48] optimizer with an initial learning rate of 0.0002. Following the completion of half of the total epochs, the learning rate linearly decayed to 0 over the remaining epochs. In TANQ, a global attention matrix was computed, and 256 features based on it were selected[18]. The dimensions of the anchor, query, and key features used in the contrastive loss were set to 256[18]. We adopted multilayer feature extraction with five layers, and the QS-Attn module was applied in the last two layers[18]. During the segmentation and self-training phases, we used a semantic segmentation method based on nnUNet[40]. The nnUNet model was trained using both Dice and cross-entropy losses, and the batch size was 8 for 1000 epochs using the SGD optimizer[40] with an initial learning rate of 0.01. The learning rate was decayed according to the 'poly' learning rate policy[49] throughout the training phase[40]. Finally, we performed an ensemble of predictions from the 2D and 3D segmentation models through the ensemble selection process within nnUNet[40].

For the image translation network, we strictly followed the default parameters and training protocols of CycleGAN[33] and QS-Attn[18], as provided in their respective GitHub repositories. The training procedures were adhered to exactly as outlined in their repositories[18,33]. We selected the final model from the last epoch, consistent with the methodology used in these baseline models[18,33]. For the segmentation network, we strictly followed the default parameters of nnUNet[40] and selected the final model using five-fold cross-validation.

Our TANQ consists of three main components, namely the Generator, Discriminator, and Feature Projector, which have 8.219M, 2.763M, and 0.560M parameters, respectively. CycleGAN[33] requires two Generators and two Discriminators for reverse mapping. Each Generator in CycleGAN[33] consists of 11.366M parameters, while each Discriminator contains 2.763M parameters. For the segmentation process, we utilize nnUNet[40], which comprises a total of 30.76 million parameters. Regarding training duration, both TANQ and CycleGAN[33] required a total of 72 GPU hours using a single RTX 3090 GPU. On the other hand, the Segmentation network required 24 GPU hours using 5 RTX 3090 GPUs. For inference time, TANQ takes approximately 0.0669 seconds per 2D image, while CycleGAN[33] performs inference in 0.0043 seconds per 2D image. For the nnUNet[40] model, inference on a 3D scan takes about 12.104 seconds.

Results and discussion

Performance evaluation

Table 1 presents the comparative results, including statistical analysis, of the proposed and competing methods for the validation set provided by the CrossMoDA 2021, 2022 challenge. Our proposed method demonstrated the best performance, with a mean Dice score of 0.8650 (± 0.0370), outperforming all the other methods. In particular, it demonstrated outstanding performance in cochlear segmentation, achieving a remarkable Dice score of 0.8750 (± 0.0217) and an ASSD of 0.1553 (± 0.1448), with statistically highly significant ($**p < 0.0001$) improvements in both metrics compared to all other methods. In the case of VS segmentation, the proposed method demonstrated the second-best performance, with a Dice score of 0.8550 (± 0.0731) and ASSD of 0.4643 (± 0.2000) after the TBA[34] method on the validation set. Additionally, our framework showed statistically

	Dice score (↑)			ASSD (↓)	
Method	**VS**	**Cochlea**	**Mean**	**VS**	**Cochlea**
COSMOS[13]	0.6104 (±0.3065)**	0.8184 (±0.0257)**	0.7144 (±0.1480)	3.8170 (±5.0257)**	0.2293 (±0.1633)**
Fgh365[21]	0.8178 (±0.0803)**	0.8433 (±0.0293)**	0.8306 (±0.0420)	0.6673 (±0.2713)**	0.2053 (±0.1489)**
MSF-Net[43]	0.8493 (±0.0683)	0.8294 (±0.0268)**	0.8394 (±0.0368)	0.5202 (±0.2288)**	0.2454 (±0.2102)**
PAST[14]	0.8473 (±0.0633)	0.8547 (±0.0283)**	0.8511 (±0.0322)	0.5513 (±0.3026)**	0.1874 (±0.1478)**
Multi-view[17]	0.8520 (±0.0889)	0.8488 (±0.0235)**	0.8504 (±0.0466)	0.4748 (±0.2072)	0.1992 (±0.1524)**
TBA[34]	**0.8682 (±0.0601)***	0.8506 (±0.0294)**	0.8594 (±0.0347)	**0.4302 (±0.1780)**	0.1892 (±0.1457)**
Ours	0.8550 (±0.0731)	**0.8750 (±0.0217)**	**0.8650 (±0.0370)**	0.4643 (±0.2000)	**0.1553 (±0.1448)**

Table 1. Comparison of segmentation results of the unsupervised domain adaptation methods from the top-ranked methods of the CrossMoDA 2021, 2022 challenge. The paired *t*-test was conducted to assess the statistical significance of our results. ($*p < 0.05, **p < 0.0001$).

highly significant $(**p < 0.0001)$ improvements in ASSD when compared to MSF-Net[43] and PAST[14], while also exhibiting significant differences $(**p < 0.0001)$ in both Dice score and ASSD against the COSMOS[13] andFgh365[21] methods. On the other hand, when compared to TBA[34], the Dice score was found to be statistically significantly lower $(*p < 0.05)$, but there was no statistically significant difference in the ASSD metric. Furthermore, our previous model (i.e., Multi-view[17]) which was adopted as the backbone framework in this study, demonstrated the best VS segmentation performance on the test dataset of the CrossMoDA 2022 challenge. As listed in Table 1, our proposed method achieved better performance in VS segmentation than our previous method in terms of both the Dice score and ASSD. Based on these results, we expected that our proposed method would also exhibit excellent performance in VS segmentation on the test dataset. The comparative results, including statistical analysis, highlight the effectiveness of considering the image characteristics of the target objects (the VS and cochlea) during image translation from ceT_1 to hrT_2 images. By normalizing these target regions separately, the proposed method enables the generation of more stable and realistic pseudo-hrT_2 images, ultimately leading to improved segmentation performance.

Ablation study results

Table 2 presents the segmentation results of the ablation study according to the individual components of the image translation method within our framework. Here, in order to directly compare the segmentation results according to the image translation methods, self-training was not applied. First, in Table2, we can observe that QS-Attn[18], which adopts entropy-based informative patch selection in contrastive learning, achieved better VS/cochlea segmentation performance than CUT[16], which conducts contrastive learning with randomly selected patches. This result implies that selecting informative patches from ceT_1 images is beneficial for preserving the structures of the VS and cochleas with contrastive learning. Second, upon comparing the results of Multi-view[17] with those of CycleGAN[33] and QS-Attn, we can assume that the multi-view image translation approach, which uses both CycleGAN and QS-Attn in parallel, leads to enhanced segmentation performance with its enriched pseudo-hrT_2 image representations. Finally, based on the proposed target-aware normalization techniques for the VS and cochleas, the proposed method further achieved enhanced segmentation performance in the ablation study. Specifically, when only RN was applied to our multi-view approach (i.e., Ours (*w/o.* RAIN)), we can observe from Table 2 that the segmentation performance of the cochleas had improved in terms of both the Dice score and ASSD. This observation suggests that RN is effective in preserving the structural information of the cochlea. Ours (*w/o.* SegDecoder) represents the method that incorporates the proposed target-aware normalization technique while excluding solely the SegDecoder. Compared to the Ours (*w/o.* RAIN), Ours (*w/o.* SegDecoder) exhibited an improvement in the VS and cochlea Dice scores. This indicates that both RN and RAIN in our target-aware normalization effectively preserve the structural information of the target objects, VS and cochlea, and accurately reflect these visual characteristics in real hrT2 images. Moreover, as shown in Table 2, Ours, which adopts both target-aware normalization and the SegDecoder, further improved the segmentation performance of VS and cochlea across all metrics compared to other methods. This result highlights Ours as a tailored approach for cross-modality image translation, specifically designed for the VS and cochlea. We also performed a statistical evaluation for our ablation study. In cochlear segmentation, Ours showed statistically highly significant differences $(**p < 0.0001)$ compared to CycleGAN[33], CUT[16], QS-Attn[18], and Multi-view[17]. For Ours (*w/o.* RAIN), statistically significant differences $(*p < 0.05)$ were observed in cochlear segmentation. In VS segmentation, Ours demonstrated statistically significant differences $(*p < 0.05)$ in both Dice score and ASSD compared to CycleGAN[33], CUT[16], and QS-Attn[18]. Notably, compared to the baseline Multi-view[17] and Ours (*w/o.* SegDecoder), a statistically significant improvement in ASSD for VS segmentation $(*p < 0.05)$ was achieved. These statistical anaylses highlight the importance of target-aware normalization during image translation when considering the different image characteristics of the target objects.

Fig. 3 exhibits the segmentation results of the real hrT_2 images obtained from segmentation models trained on psuedo-hrT_2 images from different image translation methods adopted for the ablation study. Notably, our approach excels in segmenting the VS and cochleas compared to the other models, as shown in Fig. 3, demonstrating exceptional performance by achieving significantly finer and more precise segmentation than the other methods.

Method	Dice score (↑)			ASSD (↓)	
	VS	Cochlea	Mean	VS	Cochlea
CycleGAN[33]	0.7798 (±0.1901)*	0.8066 (±0.0323)**	0.7932 (±0.0972)	0.8750 (±0.9222)*	0.2422 (±0.1608)**
CUT[16]	0.7693 (±0.2097)*	0.7950 (±0.0337)**	0.7822 (±0.1095)	0.7126 (±0.6179)*	0.2586 (±0.1668)**
QS-Attn[18]	0.7779 (±0.1825)*	0.8158 (±0.0287)**	0.7968 (±0.0929)	0.6667 (±0.3891)*	0.2365 (±0.1573)**
Multi-view[17]	0.8043 (±0.1656)	0.8158 (±0.0289)**	0.8101 (±0.0863)	0.5742 (±0.2461)*	0.2387 (±0.1581)**
Ours (*w/o.* RAIN)	0.8017 (±0.1696)	0.8372 (±0.0304)*	0.8195 (±0.0896)	1.3809 (±6.5339)	0.2070 (±0.1608)*
Ours (*w/o.* SegDecoder)	0.8057 (±0.1356)	0.8393 (±0.0299)	0.8225 (±0.0716)	1.7198 (±6.5030)*	0.2064 (±0.1597)
Ours	**0.8097 (±0.1635)**	**0.8397 (±0.0293)**	**0.8247 (±0.0863)**	**0.5708 (±0.2919)**	**0.2042 (±0.1610)**

Table 2. Quantitative results from the ablation study. The results were obtained without conducting self-training. The paired *t*-test was conducted to assess the statistical significance of our results. $(*p < 0.05, **p < 0.0001)$.

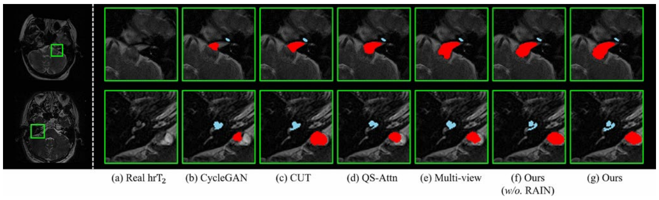

Fig. 3. Qualitative comparison of segmentation results for VS (red) and cochleas (sky-blue) on the ablation study. (**a**) is the real hrT_2 image, while (**b**) to (**g**) show the segmentation results overlaid on the real hrT_2 image (**a**). These results are obtained from segmentation models trained using pseudo-hrT_2 images generated by the corresponding translation models for (**b**) to (**g**). Note that the ground truths of the real hrT_2 scans are not accessible in the CrossMoDA challenge.

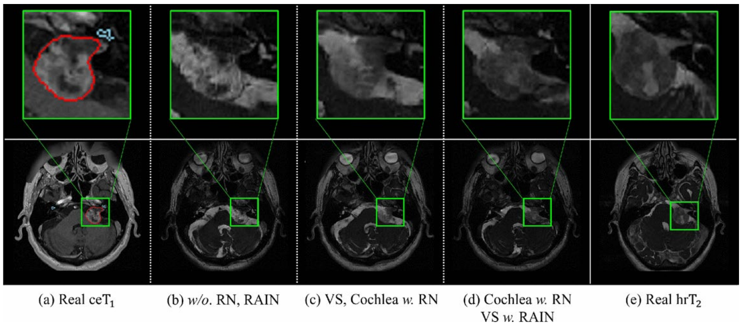
(a) Real ceT_1
(b) *w/o.* RN, RAIN
(c) VS, Cochlea *w.* RN
(d) Cochlea *w.* RN VS *w.* RAIN
(e) Real hrT_2

Fig. 4. Visualization of the effectiveness of RN and RAIN for VS (red) and cochleas (sky-blue) according to different normalization configurations of TANQ; (**a**) real ceT_1 images, (**b**)-(**d**) pseudo-hrT_2 images from ceT_1 images generated by TANQ with different normalization configurations, and (**e**) real hrT_2 images. In the real ceT_1 images, the intensity of the VS is higher than that of other tissues due to the contrast agent, while the VS in the real hrT_2 images shows similar intensity to other tissues. Conversely, the cochlea exhibits very low intensity in real ceT_1 images (**a**), whereas it appears relatively distinct in real hrT_2 images (**e**).

Effects of target-aware normalization

Visual characteristics of target objects

Fig. 4 shows the image translation results under conditions where RN[36] and RAIN[37] are included or excluded for each of the target objects. Fig. 4(a) and (e) show samples of real ceT_1 and hrT_2 images, respectively, while (b)-(d) illustrate the pseudo-hrT_2 images generated from the real ceT_1 images based on the proposed TANQ with different RN/RAIN configurations.

From the results, it can first be observed that the pseudo-hrT_2 images from (b) TANQ *w/o*. RN and RAIN represent the VS with significantly high intensity, indicating a failure to reflect the low-intensity characteristic of VS in (e) the real hrT_2 images, and the structure of the cochlea is poorly preserved and appears blurred. By contrast, the pseudo-hrT_2 images from (c) TANQ with RN applied to both the VS and cochleas better preserve the information of these structures. This implies that preserving the information of each region using RN is essential for maintaining the integrity of the VS and cochlear structures. Finally, the pseudo-hrT_2 images from (d) TANQ with RN and RAIN for cochleas and VS show a lower intensity for the VS, similar to (e) the real hrT_2 images, as well as better preserved cochlear structures. These observations suggest that the use of the RN and RAIN is essential for accurately reflecting the distinctive characteristics of the target objects and for maintaining the integrity of the VS and cochlear structures. Based on the pseudo-hrT_2 images generated with the advantages of the proposed TANQ, the segmentation model can successfully learn the realistic characteristics of the VS and cochleas from the given hrT_2 images for VS and cochlea segmentation.

Informative patch selection

The locations of the selected patches for contrastive learning in (a) CUT[16], (b) QS-Attn[18], and (c) our TANQ are visualized with green indicators in Fig.5. It can be observed from Fig. 5(a) that CUT selects patches arbitrarily. This visualization demonstrates that such random selection tends to select patches from the background area as well, rather than focusing on specific anatomical structures of the brain, resulting in the lack of crucial information regarding brain structure, including the VS and cochlea. Numerous studies have also noted that using these “easy” negative samples for contrastive learning can result in poor performance in image translation tasks[18,50,51].

In contrast, Fig. 5(b) illustrates that QS-Attn effectively calculates the contrastive loss by selecting features from regions that represent the structural information of the brain. Despite these advantages, Fig. 5(b) indicates that QS-Attn tends to struggle with feature selection in smaller VS (red) and cochlear (sky-blue) regions. In

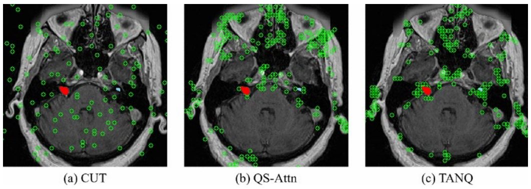

Fig. 5. Visualization of the locations(green circle) of selected features when calculating the contrastive loss. The red indicates VS, and sky-blue represents cochlea in the $\mathrm{ceT_1}$ image.

contrast, Fig. 5(c) demostrates a significant increase in the patch selection related to the regions of VS and cochlea, indicating a stronger focus on the VS and cochlea than in QS-Attn. This proves that our TANQ is remarkably effective with its target-aware normalization in preventing information loss in the VS and cochleas, guaranteeing the selection of patches specifically associated with these regions. Our framework generates more realistic pseudo-$\mathrm{hrT_2}$ images by focusing on the target objects during image translation for more efficient and effective contrastive learning techniques. Consequently, this approach enhances the performance of the segmentation model by obtaining suitable pseudo-$\mathrm{hrT_2}$ images.

Limitations and future works

Our current framework employs two independent stages: the image translation stage and the segmentation stage. This independent approach does not allow for interaction between the image translation models, such as CycleGAN[33] and TANQ, and the segmentation model, nnUNet[40]. As a result, the performance of the segmentation model is highly dependent on the quality of the image translation model. Additionally, despite the significant impact of the initial quality of the pseudo-labels on the final segmentation performance[52], our current work did not involve any refinement of the pseudo-labels.

In our future work, we will aim to develop a framework that enables bidirectional interaction between the image translation and segmentation stages. This will allow for more effective interaction between the two stages, potentially enhancing the overall performance of both tasks. Furthermore, we plan to incorporate pseudo-label refinement techniques to enhance the effectiveness of the self-training process, ultimately contributing to improved performance of the segmentation model.

Conclusion

In this study, we propose a target-aware, unsupervised domain adaptation framework for VS and cochlea segmentation. Specifically, we acquire pseudo-$\mathrm{hrT_2}$ images that reliably reflect the characteristics of real $\mathrm{hrT_2}$ images while preserving the structural information of small-sized VS and cochleas by using target-aware normalization according to the specific characteristics of the VS and cochlea. By applying this target-aware normalization to QS-Attn and integrating multi-view representation, we can construct a robust VS and cochlea segmentation model in real $\mathrm{hrT_2}$ images using these pseudo-$\mathrm{hrT_2}$ images. Our approach demonstrates its effectiveness by significantly surpassing the performance of state-of-the-art cross-modality VS and cochlea segmentation methods.

Data availability

The CrossMoDA dataset is publicly available, and more information can be found at the following link: https://crossmoda-challenge.ml/challenge2022/

Received: 6 March 2024; Accepted: 23 October 2024
Published online: 13 November 2024

References

1. Dorent, R. *et al.* Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwnannoma and cochlea segmentation. arXiv preprint arXiv:2201.02831 (2022).
2. Kentala, E. & Pyykkö, I. Clinical picture of vestibular schwannoma. *Auris Nasus Larynx* **28**, 15–22 (2001).
3. di Russo, P. et al. Characteristics and management of hydrocephalus associated with vestibular schwannomas: a systematic review. *Neurosurgical Review* **44**, 687–698 (2021).
4. Shapey, J. et al. Segmentation of vestibular schwannoma from mri, an open annotated dataset and baseline algorithm. *Scientific Data* **8**, 286 (2021).
5. McGrath, H. et al. Manual segmentation versus semi-automated segmentation for quantifying vestibular schwannoma volume on mri. *International Journal of Computer Assisted Radiology and Surgery* **15**, 1445–1455 (2020).
6. Zahara, D. et al. Variations in cochlear size of cochlear implant candidates. *International archives of otorhinolaryngology* **23**, 184–190 (2019).
7. Marasini, R., Thanh Nguyen, T. D. & Aryal, S. Integration of gadolinium in nanostructure for contrast enhanced-magnetic resonance imaging. *Wiley Interdisciplinary Reviews: Nanomedicine and Nanobiotechnology* **12**, e1580 (2020).
8. Khawaja, A. Z. et al. Revisiting the risks of mri with gadolinium based contrast agentsâ€"review of literature and guidelines. *Insights into imaging* **6**, 553–558 (2015).
9. Yao, K. et al. A novel 3d unsupervised domain adaptation framework for cross-modality medical image segmentation. *IEEE Journal of Biomedical and Health Informatics* **26**, 4976–4986 (2022).

10. Shin, H. *et al.* Sdc-uda: Volumetric unsupervised domain adaptation framework for slice-direction continuous cross-modality medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 7412–7421 (2023).
11. 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).
12. Choi, J. W. Using out-of-the-box frameworks for unpaired image translation and image segmentation for the crossmoda challenge. *arXiv e-prints* arXiv–2110 (2021).
13. Shin, H. *et al.* 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).
14. Hexin, D. *et al.* Unsupervised domain adaptation in semantic segmentation based on pixel alignment and self-training (past). CrossMoDA 2022 Challenge.
15. 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*, 8168–8177 (2020).
16. Park, T., Efros, A. A., Zhang, R. & Zhu, J.-Y. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, 319–345 (Springer, 2020).
17. Kang, B., Nam, H., Han, J.-W., Heo, K.-S. & Kam, T.-E. Multi-view cross-modality mr image translation for vestibular schwannoma and cochlea segmentation. arXiv preprint arXiv:2303.14998 (2023).
18. Hu, X. *et al.* Qs-attn: Query-selected attention for contrastive learning in i2i translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18291–18300 (2022).
19. Ulyanov, D., Vedaldi, A. & Lempitsky, V. Instance normalization: The missing ingredient for fast stylization. arXiv preprint arXiv:1607.08022 (2016).
20. Zhuang, X. & Shen, J. Multi-scale patch and multi-modality atlases for whole heart segmentation of mri. *Medical image analysis* **31**, 77–87 (2016).
21. Liu, H., Fan, Y. & Dawant, B. M. Enhancing data diversity for self-training based unsupervised cross-modality vestibular schwannoma and cochlea segmentation. arXiv preprint arXiv:2209.11879 (2022).
22. Park, T., Liu, M.-Y., Wang, T.-C. & Zhu, J.-Y. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2337–2346 (2019).
23. Dou, Q. et al. Pnp-adanet: Plug-and-play adversarial domain adaptation network at unpaired cross-modality cardiac segmentation. *IEEE Access* **7**, 99065–99076 (2019).
24. Yan, W., Wang, Y., Xia, M. & Tao, Q. Edge-guided output adaptor: Highly efficient adaptation module for cross-vendor medical image segmentation. *IEEE Signal Processing Letters* **26**, 1593–1597 (2019).
25. Chen, Z., Pan, Y. & Xia, Y. Reconstruction-driven dynamic refinement based unsupervised domain adaptation for joint optic disc and cup segmentation. *IEEE Journal of Biomedical and Health Informatics* **27**, 3537–3548 (2023).
26. Cai, J., Zhang, Z., Cui, L., Zheng, Y. & Yang, L. Towards cross-modal organ translation and segmentation: A cycle-and shape-consistent generative adversarial network. *Medical image analysis* **52**, 174–184 (2019).
27. Jue, J. *et al.* Integrating cross-modality hallucinated mri with ct to aid mediastinal lung tumor segmentation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part VI 22*, 221–229 (Springer, 2019).
28. Kang, M., Chikontwe, P., Won, D., Luna, M. & Park, S. H. Structure-preserving image translation for multi-source medical image domain adaptation. *Pattern Recognition* **144**, 109840 (2023).
29. Chen, C., Dou, Q., Chen, H., Qin, J. & Heng, P. A. Unsupervised bidirectional cross-modality adaptation via deeply synergistic image and feature alignment for medical image segmentation. *IEEE transactions on medical imaging* **39**, 2494–2505 (2020).
30. 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* **33**, 865–872 (2019).
31. Hu, S., Liao, Z. & Xia, Y. Domain specific convolution and high frequency reconstruction based unsupervised domain adaptation for medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 650–659 (Springer, 2022).
32. Jiang, J. et al. Psigan: Joint probabilistic segmentation and image distribution matching for unpaired cross-modality adaptation-based mri segmentation. *IEEE transactions on medical imaging* **39**, 4071–4084 (2020).
33. 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*, 2223–2232 (2017).
34. Sallé, G. *et al.* Cross-modal tumor segmentation using generative blending augmentation and self training. arXiv preprint arXiv:2304.01705 (2023).
35. Shaham, T. R., Dekel, T. & Michaeli, T. Singan: Learning a generative model from a single natural image. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4570–4580 (2019).
36. Yu, T. et al. Region normalization for image inpainting. *In Proceedings of the AAAI conference on artificial intelligence* **34**, 12733–12740 (2020).
37. Ling, J., Xue, H., Song, L., Xie, R. & Gu, X. Region-aware adaptive instance normalization for image harmonization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9361–9370 (2021).
38. Yi-de, M., Qing, L. & Zhi-Bai, Q. Automated image segmentation using improved pcnn model based on cross-entropy. In *Proceedings of 2004 International Symposium on Intelligent Multimedia, Video and Speech Processing, 2004.*, 743–746 (IEEE, 2004).
39. Goodfellow, I. *et al.* Generative adversarial nets. *Advances in Neural Information Processing Systems* **27** (2014).
40. 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**, 203–211 (2021).
41. 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*, 10687–10698 (2020).
42. Yang, N., Rongione, C., Jacquemart, A.-L., Draye, X. & Vleeschouwer, C. D. On the importance of diversity when training deep learning segmentation models with error-prone pseudo-labels. *Applied Sciences* **14**, 5156 (2024).
43. Han, L., Huang, Y., Tan, T. & Mann, R. Unsupervised cross-modality domain adaptation for vestibular schwannoma segmentation and koos grade prediction based on semi-supervised contrastive learning. arXiv preprint arXiv:2210.04255 (2022).
44. Dice, L. R. Measures of the amount of ecologic association between species. *Ecology* **26**, 297–302 (1945).
45. Huttenlocher, D. P., Klanderman, G. A. & Rucklidge, W. J. Comparing images using the hausdorff distance. *IEEE Transactions on pattern analysis and machine intelligence* **15**, 850–863 (1993).
46. Johnson, J., Alahi, A. & Fei-Fei, L. Perceptual losses for real-time style transfer and super-resolution. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, 694–711 (Springer, 2016).
47. Isola, P., Zhu, J.-Y., Zhou, T. & Efros, A. A. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1125–1134 (2017).
48. Kinga, D., Adam, J. B. *et al.* A method for stochastic optimization. In *International conference on learning representations (ICLR)*, vol. 5, 6 (San Diego, California;, 2015).

49. Chen, L.-C., Papandreou, G., Kokkinos, I., Murphy, K. & Yuille, A. L. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence* **40**, 834–848 (2017).
50. Jung, C., Kwon, G. & Ye, J. C. Exploring patch-wise semantic relation for contrastive learning in image-to-image translation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18260–18269 (2022).
51. Wang, W., Zhou, W., Bao, J., Chen, D. & Li, H. Instance-wise hard negative example generation for contrastive learning in unpaired image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 14020–14029 (2021).
52. Higashimoto, R., Yoshida, S., Horihata, T. & Muneyasu, M. Unbiased pseudo-labeling for learning with noisy labels. *IEICE TRANSACTIONS on Information and Systems* **107**, 44–48 (2024).

Acknowledgements

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant (No. 2019-0-00079, the Artificial Intelligence Graduate School Program (Korea University)) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS202300212498). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2021-II212068, Artificial Intelligence Innovation Hub).

Author contributions

B.K. and T.-E.K. conceived the study and designed the methodology. B.K. implemented the source code and performed the related experiments. B.K. and T.-E.K. conducted the analysis. H.N. supported B.K. in drafting the manuscript under the supervision of T.-E.K. M.K supported B.K. in designing the methodology. H.N., M.K., K.-S.H., M.L., and J.-H.O. reviewed and edited the manuscript under the supervision of T.-E.K.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to T.-E.K.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.