IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 15 March 2024, accepted 19 May 2024, date of publication 22 May 2024, date of current version 30 May 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3404071

RESEARCH ARTICLE

US-GAN: Ultrasound Image-Specific Feature Decomposition for Fine Texture Transfer

SEONGHO KIM, (Associate Member, IEEE),
AND BYUNG CHEOL SONG, (Senior Member, IEEE)
Department of Electrical and Computer Engineering, Inha University, Incheon 22212, Republic of Korea

Corresponding author: Byung Cheol Song (bcsong@inha.ac.kr)

This work was supported in part by the Institute of Information and Communications Technology Planning and Evaluation (IITP) grant funded by Korean Government (MSIT) [Artificial Intelligence Convergence Innovation Human Resources Development (Inha University)] under Grant RS-2022-00155915 and (AI Innovation Hub) Grant RS-2021-II212068, and in part by the National Research Foundation of Korea (NRF) grant funded by Korean Government (MSIT) under Grant 2022R1A2C2010095 and Grant 2022R1A4A1033549.

ABSTRACT Ultrasound images acquired through various measuring devices may have different styles, and each style may be specialized for diagnosing specific diseases. Accordingly, ultrasound image-to-image translation (US I2I) has become an essential research field. However, direct application of conventional I2I techniques to US I2I is difficult because it causes content deformation and has the problem of not being able to accurately translate fine textures. To solve the aforementioned problems, this paper proposes a novel feature decomposition scheme specialized for US I2I. The proposed feature decomposition explicitly separates texture and content information in latent space. Then, fine textures of the US image are effectively translated through translation of only the texture features. Moreover, I2I is carried out in a way that minimizes changes to the original content through reuse of content features. In addition to the feature decomposition scheme, we present a contrastive loss designed for content preservation. Specifically, the contrastive loss can maximize the content preservation effect because it preferentially performs query selection, which allows regions containing organ structures to be selected as queries (i.e., anchors). The proposed US image-specific learning scheme leads to qualitatively superior results, and the excellence of each method has been experimentally verified through various quantitative metrics.

INDEX TERMS Unpaired image-to-image tranlsation, ultrasound image, feature decomposition, contrastive learning.

I. INTRODUCTION

Recently, with the development of neural networks, the growth of the ultrasound (US) image processing field is accelerating. In particular, the achievements in deep learning-based US image processing, such as disease classification and lesion segmentation, are remarkable [1], [2], [3], [4], [5], [6]. Such technological progress is expanding to more high-level tasks, e.g., the field of image generation [7], [8], [9].

US images can have various styles depending on the acquisition method or equipment, and there is no absolute "correct style." Because of these characteristics, US image-to-image translation (shortly, US I2I) is emerging as an important research topic in the field of deep learning-based medical image processing. For example, we can imagine a scenario where the US image of a specific device is changed to the US image style of the preferred device depending on the doctor's preference. However, US I2I is quite challenging because the technical difficulty of acquiring different styles of US images for the same scene is very high.

Meanwhile, unpaired I2I methods can be effective for US images that are difficult to configure in pairs. Note that the core of unpaired I2I is to preserve the content (e.g., shape) of the image and translate its appearance (e.g., texture). For this purpose, Zhu et al. [10] proposed the so-called cycle-consistency loss, which measures the pixel distance between the input image and the reconstructed image. However,

The associate editor coordinating the review of this manuscript and approving it for publication was Chao Zuo.

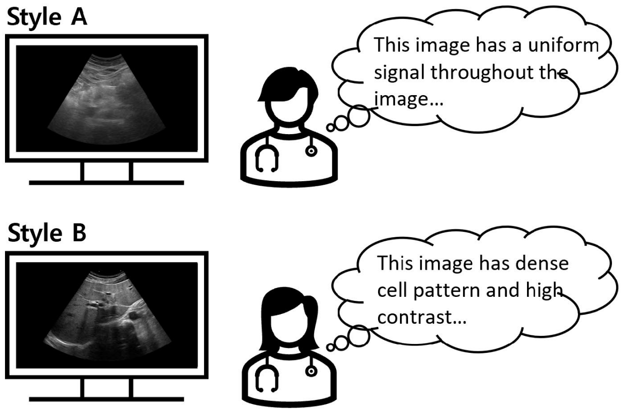

Style A
This image has a uniform signal throughout the image...
Style B
This image has dense cell pattern and high contrast...

FIGURE 1. Description of each US image. Each US image has its own strengths.

cycle-consistency loss may increase training costs because it requires two generators. In addition, cycle-consistency loss has limitations in preserving local details because it receives the entire image as input. To solve these problems, CUT [11] adopted patch-wise contrastive learning, which can maximize content representation between the same spatial regions with only a single generator. This feature representation scheme is learned through contrastive learning between latent features of the source and translated images. Thus, CUT effectively reduces training costs and improves generation performance (i.e., content preservation) even in situations where there is no clear ground truth (GT). However, conventional methods still have difficulty in completely preserving the content and translating the texture, and only focus on natural images. Therefore, it is realistically hard to directly apply them to US I2I task, which requires translation of fine texture while preserving organ structure.

Specifically, let's look at the limitations of conventional techniques in terms of content preservation and texture translation. Generally, conventional techniques map the input image to latent space through an encoder and then attempt translation to the target domain. At this time, the attributes (i.e., content and texture) are entangled in the encoded latent feature [12], [13]. So, if we decode the latent features directly into the target domain, not only relevant attributes (i.e., texture) can be translated, but even irrelevant attributes (i.e., content) can be changed [14], [15]. Such a change can lead to damage to fine organ structures during the US I2I process, and it is difficult to translate targeting only relevant attributes, resulting in qualitative limitations in terms of texture expression. Therefore, this paper defines the ultimate goal of US I2I as disentangling only relevant attributes and mapping them to the target domain.

We present US image-specific generative adversarial networks (US-GAN) that can effectively translate fine textures while preserving the various contents of US images. In detail, we propose feature decomposition that can explicitly separate content and texture in latent space. The proposed feature decomposition scheme normalizes modules through representation learning that takes into account the consistent characteristics (i.e., similar texture) that only US images have. As a result, effective US I2I is realized through translation targeting only texture features and reuse of pre-separated content features.

Additionally, we combine patch-wise contrastive learning [11] with a query selection scheme that can further improve the effect of content preservation [16]. Note that a region with a prominent local structure needs to be selected as a query in that it provides information that the query must preserve for patch-wise contrastive learning. For this purpose, we measure self-similarity between patches within the source latent feature and select the latent feature in the region containing the local structure as a query. This query selection scheme, along with the feature decomposition scheme mentioned above, shows amazing preservation performance of detailed structures.

Contributions of this paper are as follows:

- We propose a framework specialized for US I2I and a learning scheme for it. The proposed method guarantees better qualitative and quantitative performance than existing techniques.
- We propose a learning scheme that can explicitly decompose content and texture features in latent space through contrastive learning that takes into account the consistent characteristics of US images. Since the proposed feature decomposition scheme was designed considering unique pattern(s) that US images in a mini-batch have in common, it not only successfully translates fine textures but also preserves the contents well.
- To preserve the detail structure contained in US images, we propose a novel query selection scheme that can improve the effectiveness of existing patch-wise contrastive learning. It was experimentally proven that the combination of patch-wise contrastive learning and query selection scheme is effective in terms of preserving detail structures.

II. RELATED WORKS

A. IMAGE-TO-IMAGE TRANSLATION

Image-to-image translation (I2I) is classified into paired I2I and unpaired I2I according to data structure. First, paired I2I refers to a task for which GT is clearly defined when mapping ($\mathcal{X} \rightarrow \mathcal{Y}$) from source domain $\mathcal{X}$ to target domain $\mathcal{Y}$. Pix2pix [17] is the first case of realizing paired I2I by adding an L1 regularization term between translated image and GT to the existing GAN loss. In addition, Pix2pix-HD [18] overcame the limitation of not being able to express detailed texture in high resolution image translation by using two generators, i.e., global generator and local generator. Furthermore, Kim and Cho [19] alleviated the problem of existing techniques in which boundaries are translated unclearly through comparison between high frequencies in the frequency domain. However, since matching data pairs between domains is a difficult task in most configurations, I2I for unpaired configurations can be regarded as more desirable.

Early unpaired I2I [10], [20] focused on pixel level constraint using reconstructed images. They learn $\mathcal{X} \underset{G_1}{\rightarrow} \mathcal{Y}$ and $\mathcal{Y} \underset{G_2}{\rightarrow} \mathcal{X}$ using two generators G_1 and G_2. And the process of $\mathcal{X} \underset{G_1}{\rightarrow} \mathcal{Y} \underset{G_2}{\rightarrow} \mathcal{X}'$ gives the model a regularization effect for content preservation by measuring the pixel-wise distance between $\mathcal{X}$ and $\mathcal{X}'$. However, this approach not only increases training costs, but also has the disadvantage that the pixel-wise distance does not reflect the spatial structure of a specific area. To alleviate this problem, Park et al. [11] first applied patch-wise contrastive learning to an unpaired I2I task. This patch-wise contrastive learning aimed to preserve the same structural information between input and translated images in pixel space by using the same channel instances of latent features. This learning scheme inspired subsequent studies. For example, F-LSeSim [21] used an auxiliary encoder to extract regions with high self-similarity and performed patch-wise contrastive learning using them. Furthermore, in the most recent study, i.e., Qs-attn [16], a model was designed to measure self-similarity using only the existing encoder, and a learning scheme in which regions containing important structural information are selected as queries by introducing an attention mechanism was proposed. It is true that the previous studies have alleviated existing problems to some extent. However, they still show limitations in both texture translation and content preservation because they do not consider attribute entanglement issues.

B. DISENTANGLED REPRESENTATION IN THE LATENT SPACE

The goal of disentangled representation learning is to identify and generate explanatory factors in latent space. This learning scheme achieves feature disentanglement within the model through a parametric module or at the loss stage. In particular, in the latter case, many vision tasks were solved using various metrics (e.g., Wasserstein, mean discrepancy etc.) or self-supervised learning schemes [22], [23]. In addition, the video-driven generation task focuses on the consistent characteristics of the input data to generate positive and negative sample(s) and achieves the above-mentioned goal through contrastive learning using them. For example, Behramann et al. [24] encoded stationary features (i.e., subject's identity) and non-stationary features (e.g., motion) separately, and achieved successful disentangled representation through self-supervised learning using the two features. Furthermore, Tulyakov et al. [25] used two discriminators (video discriminator and image discriminator) to explicitly separate the face identity component that appears the same for each frame from the facial expression component that appears differently, and then proposed a method to manipulate them independently. Inspired by this learning scheme, we devise a module that can explicitly separate the attributes (i.e., texture) that US images in a mini-batch typically have and the objective for learning the module. Finally, we design an attribute-wise disentangled representation based on them.

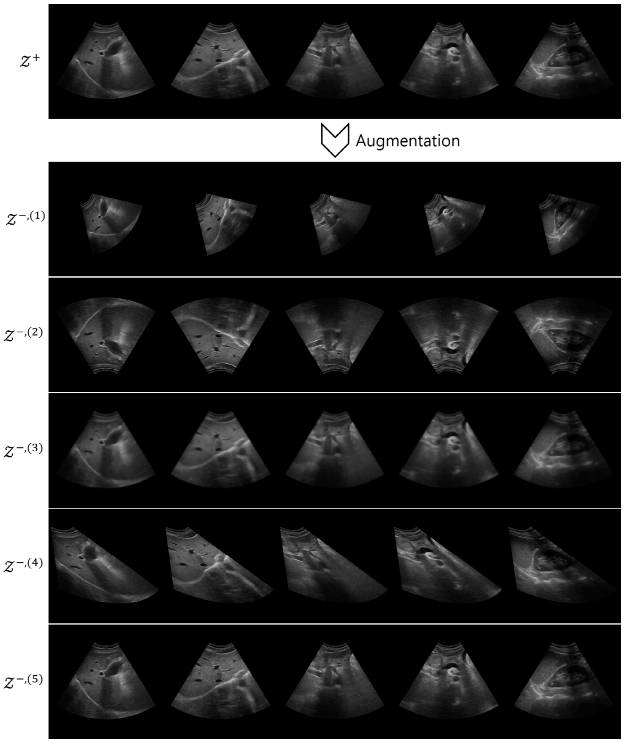


FIGURE 2. Examples of positive and negative samples. $\mathcal{Z}^{+}$ is a positive sample and $\mathcal{Z}^{-,(n)}$ is a negative sample. Here, *n* is the index of augmentation method: 1) rotation, 2) vertical flip, 3) Gaussian blur, 4) perspective transformation, and 5) adding speckle noise. Please zoom in on the figure to check the Gaussian blur and speckle noise.

III. PRELIMINARIES

This section defines problem(s) arising from conventional I2I and presents an objective for performing feature decomposition in latent space.

A. PROBLEM DEFINITION

This paper aims to translate source style image $\boldsymbol{x}$ to target style image $\tilde{x}$ through generator G. The detailed translation process is described as follows: $z = E(\boldsymbol{x}) \rightarrow \tilde{\boldsymbol{x}} = D(z)$, where E and D are the encoder and decoder of G, respectively. Here, the latent feature z of $\boldsymbol{x}$ can be subdivided into texture feature z^t and content feature z^c ($\because z := \left\{z^t, z^c\right\}$). Therefore, directly translating through $\tilde{\boldsymbol{x}} = D(z)$ may cause the following problem:

Problem 1. *When $\tilde{\boldsymbol{x}}$ is generated through $D(z)$, z^t as well as z^c are affected. So, transformation may occur in both features. In other words, the feature representation that should be required in the translated image $\tilde{\boldsymbol{x}}$ is $\left\{\tilde{z}^t, \underline{z^c}\right\}$, but the image corresponding to $\left\{\tilde{z}^t, \underline{\tilde{z}^c}\right\}$ can be produced.*

According to Problem 1, a model that fails to target only z^t not only cannot translate the fine textures, but also cannot preserve content because even z^c is manipulated. Therefore, if z^t and z^c are explicitly separated in latent space and only z^t is translated through an independent translation module, we will be able to translate up to fine textures while preserving the content. Thus, we pursue this strategy.

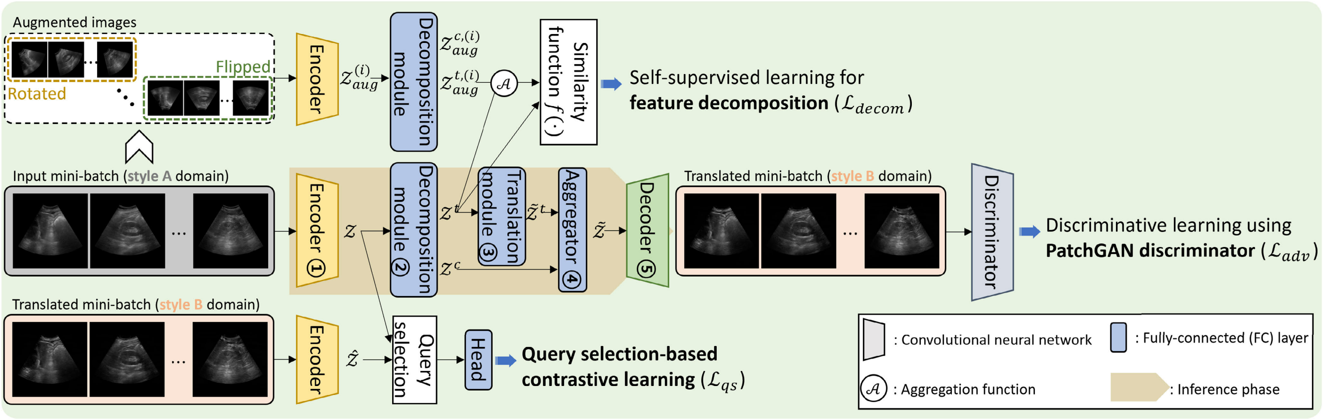

Augmented images
Rotated
Flipped
Encoder
$\mathcal{Z}^{(i)}_{aug}$
Decomposition module
$\mathcal{Z}^{c,(i)}_{aug}$
$\mathcal{Z}^{t,(i)}_{aug}$
$\mathcal{A}$
Similarity function $f(\cdot)$
Self-supervised learning for feature decomposition ($\mathcal{L}_{decom}$)
Input mini-batch (style A domain)
Encoder ①
z
Decomposition module ②
z^t
Translation module ③
$\tilde{z}^t$
Aggregator ④
$\tilde{z}$
Decoder ⑤
z^c
Translated mini-batch (style B domain)
Discriminator
Discriminative learning using PatchGAN discriminator ($\mathcal{L}_{adv}$)
Translated mini-batch (style B domain)
Encoder
$\hat{z}$
Query selection
Head
Query selection-based contrastive learning ($\mathcal{L}_{qs}$)
: Convolutional neural network
: Fully-connected (FC) layer
$\mathcal{A}$: Aggregation function
: Inference phase

FIGURE 3. **Overview of the proposed US-GAN: The weights of the encoder and decomposition modules of each path are shared. In the testing phase, only the central path is used.**

B. DEFINING THE OBJECTIVE FOR TEXTURE-AWARE DECOMPOSITION

We employ contrastive loss [26], which is often used in self-supervised learning, for explicit feature separation in latent space. The contrastive loss functions to regularize the model based on the similarity between the original feature (i.e., anchor) and positive/negative example(s). Because the consistent characteristics of domain $\mathcal{X}$ are reflected in the original feature set $\mathcal{Z}$ of input image X composed of mini-batch units, its aggregation $\mathcal{Z}^+$ can be considered a positive example. Additionally, augmentation techniques (e.g., Gaussian blur) that can cause visible transformation of the target attribute form a distribution that is different from the original. Thus, the aggregation $\mathcal{Z}^-$ of features extracted from augmented images can be considered a negative example [27] (see Fig. 2). This learning method has already been proven to be good at distribution learning in self-supervised learning without labels [25], [28]. Therefore, a contrastive objective using anchor z ($\in \mathcal{Z}$) and candidate set $\mathcal{N} := \left\{\mathcal{Z}^{-,(n)} | n \in \{1, \cdots, N\}\right\} \cup \left\{\mathcal{Z}^+\right\}$ is designed as follows:

$$\sup_{f\in\mathcal{F}} \mathbb{E}_{\mathcal{N}}\left[\log\left(\frac{\exp(f(z,\mathcal{Z}^+))}{\sum_{\bar{z}\in\mathcal{N}}\exp(f(z,\bar{z}))}\right)\right], \tag{1}$$

where f is a similarity function such as cosine similarity. Therefore, by maximizing the similiarity of z and $\mathcal{Z}^+$, the model can learn the characteristics of the target attribute. We aim to achieve successful feature decomposition by using only z^t as input to the contrastive objective.

IV. PROPOSED METHODS

As seen in the inference path of Figure 3, the proposed ultrasound image-to-image translation (shortly, US I2I) consists of the following five processes: **1)** Extract latent feature $z(\in \mathcal{Z})$ of input image x through encoder E, and **2)** separate z into $\left\{z^t, z^c\right\}$ using the feature decomposition module, and **3)** perform $z^t \rightarrow \tilde{z}^t$ using the translation module, and **4)** integrate $\left\{\tilde{z}^t, z^c\right\}$ into $\tilde{z}$ using the aggregator, and **5)** generate target style image $\tilde{x}$ using decoder D. Here, E and D are convolutional neural networks (CNNs), the decomposition and translation module is an MLP consisting of a linear layer and LeakyReLU, and the aggregator is a linear layer. In addition, since there is mapping to the target domain before decoding using a translation module, D is intended for simple up-scaling rather than generating a target domain image through transposed convolution. The next sections analyze in detail the learning schemes to realize fine texture translation and content preservation through the proposed US I2I framework.

A. SELF-SUPERVISED LEARNING FOR FEATURE DECOMPOSITION

The proposed feature decomposition module is designed to explicitly separate consistent characteristics (i.e., texture) and inconsistent characteristics (i.e., content) of US images (see the top path of Figure 3 for more visual details). To this end, the feature decomposition module is optimized by contrastive loss taking as input latent features of the input mini-batch X ($\subset \mathcal{X}$) and augmented images X_{aug}.

First, augmented images for producing negative examples are generated using the following five functions: Rotation, vertical flip, Gaussian blur, perspective transformation, and adding speckle noise (see the examples of augmented images in Figure 2). And, the features for X and X_{aug} are extracted through E as follows (① of Figure 3):

$$\mathcal{Z} = E(X), \quad \mathcal{Z}^{(i)}_{aug} = E(X^{(i)}_{aug}), \quad i \in \{1, \cdots, I\}, \tag{2}$$

where i indicates the index for each augmentation function, and I is 5. Then, the decomposition module Δ separates $\mathcal{Z}$ and $\mathcal{Z}_{aug}$ into texture and content features (refer to ② of Fig. 3):

$$\Delta(\mathcal{Z}) = \left\{\mathcal{Z}^t, \mathcal{Z}^c\right\}, \quad \Delta(\mathcal{Z}^{(i)}_{aug}) = \left\{\mathcal{Z}^{t,(i)}_{aug}, \mathcal{Z}^{c,(i)}_{aug}\right\}, \tag{3}$$

where $\mathcal{Z}^t$ and $\mathcal{Z}^c$ refer to texture and content features, respectively. The consistent features of $\mathcal{Z}^t$ and $\mathcal{Z}^{t,(i)}$ separated by

Eq. 3 are represented through the following process:

$$\Psi^t = \mathcal{A}(\mathcal{Z}^t), \quad \overline{\Psi}^{t,(i)} = \mathcal{A}(\mathcal{Z}^{t,(i)}_{aug}), \tag{4}$$

where $\mathcal{A}(\cdot)$ refers to the aggregation function, and *sum* was used as the aggregation function. According to the interpretation of Sec. III-B, Ψ^t and $\overline{\Psi}^{t,(i)}$ by Eq. 4 can be regarded as positive and negative examples respectively in terms of texture. Finally, we optimize the module using Eq. 1 designed in Sec. III-B for feature decomposition:

$$\mathcal{L}_{decom} = -\mathbb{E}_{\mathcal{N}}\left[\log\left(\frac{\exp(f(z^t, \Psi^t))}{\sum_{\overline{\Psi}^t \in \mathcal{N}} \exp(f(z^t, \overline{\Psi}^t))}\right)\right], \tag{5}$$

where z^t is a feature randomly sampled from $\mathcal{Z}^t$, and $\mathcal{N}$ is $\left\{\overline{\Psi}^{t,(i)} | i \in \{1, \cdots, I\}\right\} \cup \left\{\Psi^t\right\}$. Also, $f(\cdot)$ means cosine similarity. Therefore, according to Eq. 5, the feature decomposition module can learn attribute-wise disentangled representation in latent space. Since translation is performed by targeting only z^t, it becomes possible to express fine textures of US images in pixel space. Additionally, thanks to the reuse of z^c, the content features of the input image can be preserved, which helps preserve organ structures at the pixel level. A more detailed explanation will be available in Sec. V-D.

B. QUERY SELECTION-BASED CONTRASTIVE LEARNING

With explicit separation of z^c, we introduce the attention-based learning scheme of the latest research, i.e., Qs-att [16], as a regularization term to maximize the effect of content preservation (see the bottom path of Figure 3 for more visual details). The objective of Qs-attn was designed based on CUT [11]. The basic loss function composition is as follows:

$$\mathcal{L}_{qs} = -\log\left[\frac{\exp(q \cdot k^+/\tau)}{\exp(q \cdot k^+/\tau) + \sum_{i=1}^{N-1} \exp(q \cdot k^{-,(i)}/\tau)}\right], \tag{6}$$

where q, k^+, and $k^{-,(i)}$ mean query, positive, and negative, respectively. Among them, q is a standard factor for content preservation, and Qs-attn focuses on the selection of q. If the spatial region corresponding to the background of the US image is selected as q, the function of the target objective (i.e., content preservation) cannot operate correctly. Therefore, we try to boost the effect of content preservation by adopting a q selection scheme based on self-attention [29] and information theory (i.e., shannon entropy) [30].

Revisiting for Qs-Attn Objective: q selection aims to define a quantitative value that reflects the importance of features per spatial region. To this end, query $\boldsymbol{Q}$, key $\boldsymbol{K}$, and value $\boldsymbol{V}$ are defined by latent feature $z \in \mathbb{R}^{H\times W\times C}$ for $\boldsymbol{x}$ (by using ① in the Fig. 3) (Here, $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V} \in \mathbb{R}^{HW\times C}$). Then, the attention matrix is obtained by the following equation:

$$\boldsymbol{A} = \mathrm{softmax}(\boldsymbol{Q} \cdot \boldsymbol{K}), \quad \boldsymbol{A} \in \mathbb{R}^{HW\times HW}, \tag{7}$$

where softmax$(\cdot)$ is calculated in the row direction. Here, each row of $\boldsymbol{A}$ contains information of the same location in the spatial region because it was measured with instances of the same channel. Then, importance for each row is computed according to the following equation:

$$\boldsymbol{H}(i) = -\sum_{j=1}^{HW} \boldsymbol{A}(i,j) \log \boldsymbol{A}(i,j), \tag{8}$$

where i and j refer to the row and column of $\boldsymbol{A}$, respectively. According to Eq. 8, the closer $\boldsymbol{H}(i)$ is to 0, the fewer K positions in the i-th row are judged to be similar to the ith Q. So, by sorting $\boldsymbol{H}(i)$ in ascending order and selecting the smallest N rows as q-selection matrix $\boldsymbol{A}_{qs} \in \mathbb{R}^{N\times HW}$, query candidates and keys (i.e., positive and negatives) can be chosen within the texture region. As a result, query candidates and keys (i.e., positive and negatives) are selected within the area with the texture. Finally, the input of $\mathcal{L}_{qs}$, i.e., $\boldsymbol{q}$ and $\boldsymbol{k} := \left\{\boldsymbol{k}^+\right\} \cup \left\{\boldsymbol{k}^{-,(i)} | i \in \{1, \cdots, N-1\}\right\}$ are defined as follows:

$$\boldsymbol{q} = head(\boldsymbol{A}_{qs} \cdot \hat{\boldsymbol{V}}), \quad \boldsymbol{k} = head(\boldsymbol{A}_{qs} \cdot \boldsymbol{V}), \tag{9}$$

where since $\boldsymbol{q}$ must be selected within the translated image, the translated value obtained through $\hat{z} \in \mathbb{R}^{H\times W\times C}$, the latent feature, is used. Additionally, $head(\cdot)$ plays a role in mapping $\boldsymbol{q}$ and $\boldsymbol{k}$ to a low-dimensional space for effective contrastive learning. Further, we use $\hat{z} = E(\tilde{x})$ instead of the output of Δ, i.e., $\tilde{z}$ (after ④ in Fig. 3) for regularization of D (⑤ in Fig. 3) (see Fig. 3, ⑤ → ①).

C. OVERALL OBJECTIVES AND TRAINING PROCEDURE

The final objective function is defined as follows:

$$\mathcal{L} = \mathcal{L}_{adv} + \lambda_{decom}\mathcal{L}_{decom} + \lambda_{qs}\mathcal{L}_{qs}, \tag{10}$$

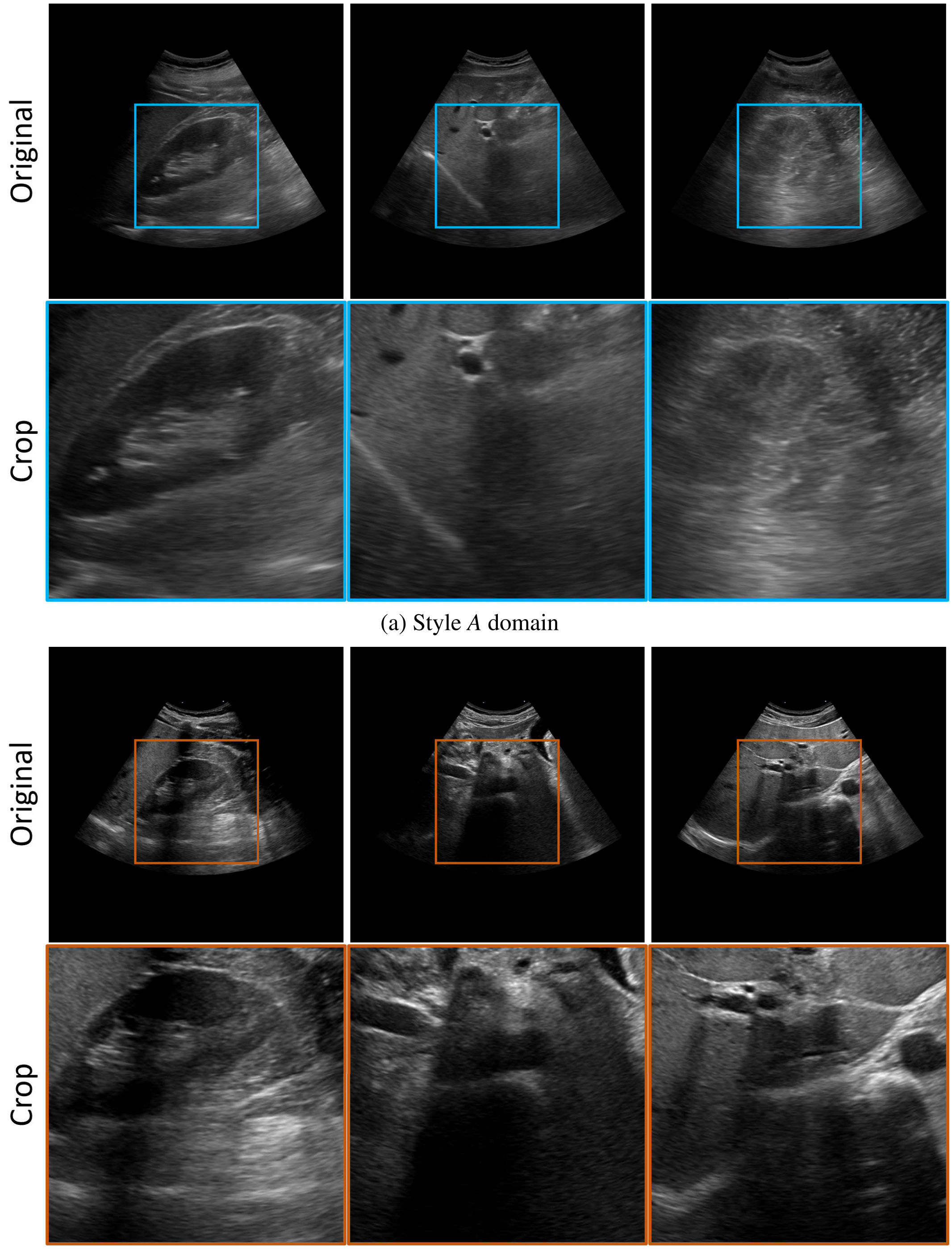
where $\mathcal{L}_{adv}$ is the loss function of LSGAN [31] for discriminative learning. In this paper, $\mathcal{L}_{adv}$ is designed using the PatchGAN discriminator [32], which is effective in fine texture translation. Note that $\mathcal{L}_{adv}$ is the main loss function for texture translation, the target objective of I2I. Meanwhile, since $\mathcal{L}_{decom}$ provides the model with the opportunity to learn disentangled representations by self-supervising disentangled features (i.e., texture features, $\mathcal{Z}^t$ and $\mathcal{Z}^t_{aug}$), it can boost the effectiveness of $\mathcal{L}_{adv}$. Therefore, Eq. 10 is an objective function specialized for fine texture translation and content preservation.

V. EXPERIMENTS

A. DATASETS AND CONFIGURATIONS

We use two datasets to learn US I2I over two domains. Two datasets consisted of 3000 images as a training set and 800 images as a validation set. Additionally, the original resolution of the images was 1024 × 1024, and was resized to 256 × 256 for the experiment.

- **Style A domain** are abdominal ultrasound (US) images acquired by an equipment of a specific A company (see Fig. 4 (a)). The advantage of style A domain is that it has



(a) Style *A* domain

(b) Style *B* domain

FIGURE 4. Visual reference of each style domain. Cropped images clearly represent the characteristics of each domain.

an overall uniform signal. However, images in this domain have the disadvantage of low lateral resolution. Here, lateral resolution refers to the minimum distance that can separately distinguish two adjacent structures (cells in US images) in the horizontal direction within an image. Therefore, the lower the lateral resolution, the longer the cell length, which makes it impossible to clearly separate two adjacent organ structures.

- **Style *B* domain** are abdominal US images acquired from equipment of a specific *B* company (see Fig. 4 (b)). Style *B* domain has the advantage of high lateral resolution. Because the cell length is short, the distinction between adjacent structures is clear. Additionally, images in the *B* domain have high contrast. This is an advantage in terms of contour detection, which can improve the accuracy of

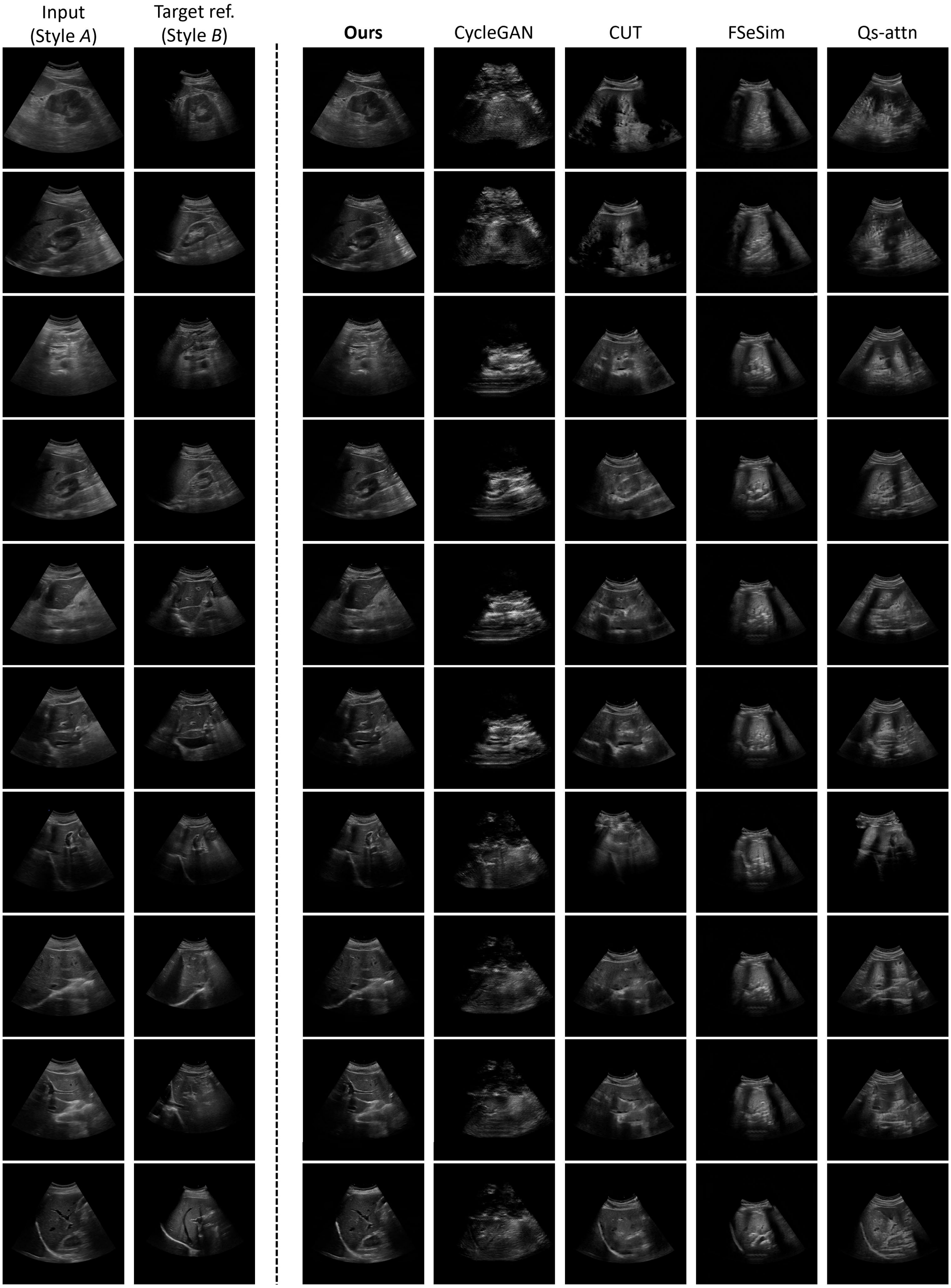


FIGURE 5. Visual quality comparison with the other methods. For each method, the input image (style *A*) is translated into the style *B* domain. Target ref. refers to the reference image of the style *B* domain and was used for texture comparison with translated images.

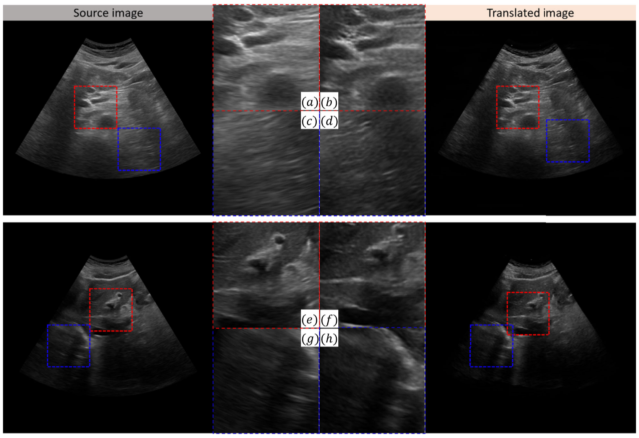


FIGURE 6. Detailed qualitative results for the proposed method. Red dotted boxes (i.e., (*a*), (*b*), (*e*), and (*f*)) represent regions containing content (particularly complex organ structures), while blue dotted boxes (i.e., (*c*), (*d*), (*g*), and (*h*)) represent regions where texture is expressed in bulk.

TABLE 1. Quantitative comparison with other methods. Bold and underlined indicate the first and second place performances, respectively.

Methods	FID (↓)	SSIM (↑)	RMSC (↑)
CycleGAN [10]	203	0.03	24.9
CUT [11]	77	0.12	24.4
FSeSim [21]	70	0.24	24.1
Qs-attn [16]	69	0.60	25.9
US-GAN (ours)	**52**	**0.73**	**34.4**

diagnosis. However, it has the disadvantage of having sparse signals than the *A* domain.

Additionally, we utilize the BraTS dataset [33] to measure the generalization performance of the proposed method. The training dataset and test dataset consist of 8457 and 979 images, respectively. This dataset includes T1 and T2 MRI images, and we aim to perform translation from T1 to T2. The images are provided at a resolution of 256 × 256.

Implementation Details. PyTorch library [34] was used to implement the proposed model. Computing for learning was done on AMD EPYC 7413 CPU and NVIDIA RTX A6000 GPU. The parameters of convolutional and FC layers were updated through Adam optimizer [35] with learning rate (LR) 10^{-4}. Additionally, the Adam optimizer performs weight decay through L2 regularization of 10^{-4}. The mini-batch size used for learning was set to 8. Finally, τ in Eq. 6 is 0.07 and λ_{decom} and λ_{qs} in Eq. 10 were set to 0.5 and 10, respectively.

B. EVALUATION SETTINGS

1) BASELINE METHODS

We compare the proposed method with the following four techniques to prove its superior quantitative and qualitative performance: CycleGAN [10], CUT [11], F-LSeSim [21], and Qs-attn [16]. Here, F-LSeSim is FSeSim trained with only a single encoder, and only the *Global* attention method is used in Qs-attn.

2) EVALUATION METRICS

To evaluate the similarity between target images and translated images, Fréchet Identity Distance (FID) [36] was used. Since FID is measured using the Inception-V3 [37] model pre-trained with ImageNet [38], we use images expanded to 3 channels (RGB). In addition, Structural Similarity Index Map (SSIM) was adopted to quantify content preservation between target images and translated images. SSIM has a range of $[-1, 1]$, and the closer it is to 1, the better the structure (i.e., content) preservation performance. Note that the calculation of SSIM involves not only the structure of the image but also luminance and contrast. So, we conduct experiments assuming the situation of $StyleB \xrightarrow[E,D]{} StyleB$ to ensure a fair comparison of techniques solely in terms of content preservation. Lastly, Root Mean Square Contrast (RMSC) [39] was used to measure the contrast of the region where the texture is distributed (i.e., inside the convex) in the US image. RMSC is defined by $\sqrt{\frac{1}{MN}\sum_{i=1}^{M}\sum_{j=1}^{N}(L_{ij}-\bar{L})^2}$. Here, L_{ij} is the pixel intensity, $\bar{L}$ is the average intensity of the image, and M x N are the image size. Since the convex size cannot be defined in terms of M and N, we use *CenterCrop* images with a size of 128×128 in the experiment. We assume that RMSC will allow us to determine the general contrast level inside the convex.

C. VERIFICATION OF THE PROPOSED METHOD

This section compares the qualitative and quantitative performance of the proposed method and baselines.

1) QUALITATIVE RESULTS

Fig. 5 shows the results of translating input images to the target domain (style *B*) using each technique. We can observe that the proposed US-GAN well expressed the texture of the *B* domain while preserving the content of input images in US I2I, which requires fine texture translation, a relatively high-level task. On the other hand, compared to US-GAN, other techniques show qualitatively poor results (see Fig. 6). Also, US-GAN maintains the shape well when regions containing complex organ structures are translated (see $(a)\rightarrow(b)$ and $(e)\rightarrow(f)$ in Fig. 6). Additionally, from (d) and (h) in Fig. 6, we can see that the proposed method can translate the texture of the target domain in terms of lateral resolution.

2) QUANTITATIVE RESULTS

Table 1 shows that US-GAN shows excellent performance in terms of fine texture translation and content preservation. Based on FID, US-GAN shows a gap of up to 151 compared to other techniques. This supports the fact that US-GAN is superior to other techniques in terms of texture expression of domain *B* in most outputs. Additionally, US-GAN shows a high SSIM of 0.73. US-GAN has outstanding content preservation performance in that SSIM is used as a measure of structural preservation of the original image. This is very important due to the nature of US I2I, in which organ structures must not change. In particular, the validity

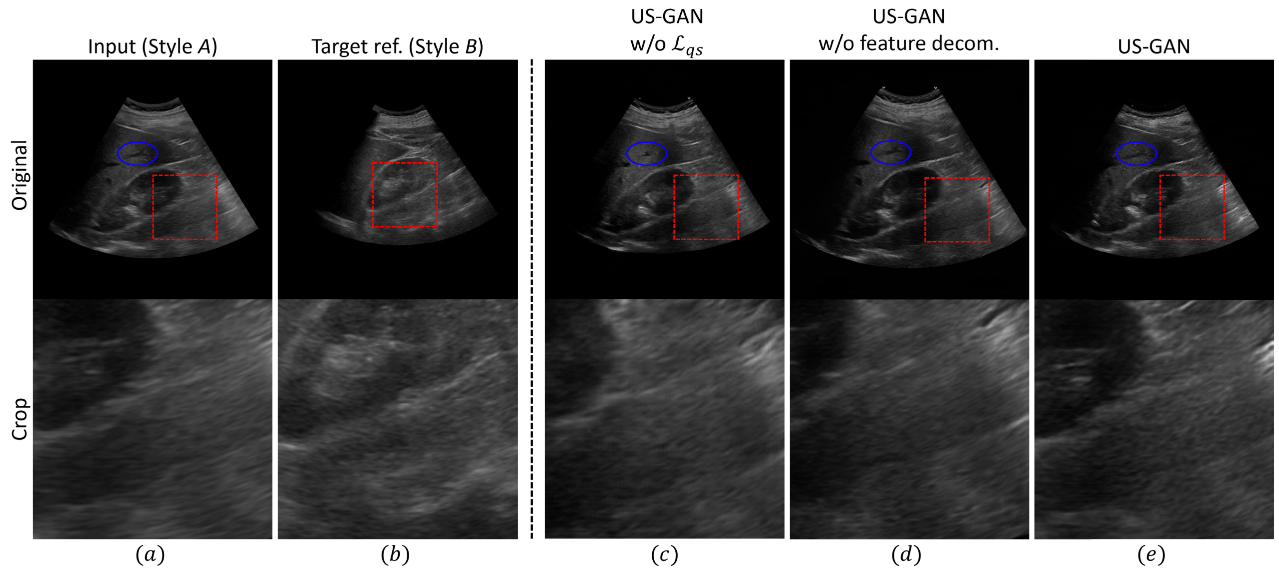


FIGURE 7. Qualitative comparison according to the application of each learning scheme. The structures represented in the blue circle are vessels. And the red dotted box indicates the part where the texture is expressed the most.

TABLE 2. Ablation study results. Except for the decomposition method and losses, other configurations remained the same. Here, baseline refers to a model trained with only "*discriminative learning using PatchGAN discriminator*" in Fig. 2. Bold and underlined indicate the first and second place performances, respectively.

Methods	Decom.	$\mathcal{L}_{qs}$	FID (↓)	SSIM (↑)	RMSC (↑)
Baseline			128	0.15	21.9
US-GAN		✓	72	<u>0.62</u>	25.7
	✓		<u>63</u>	0.55	<u>28.7</u>
	✓	✓	**52**	**0.73**	**34.4**

of selecting $\mathcal{L}_{qs}$ as the regularization term for content preservation is verified because the SSIM of Qs-attn is only 0.60. Lastly, the RMSC of the proposed method was 34.3, which is up to 10.3 higher than that of other techniques. This is interpreted as the output images of US-GAN having overall high contrast, which can be regarded as the excellent texture translation performance of the proposed method.

D. ABLATION STUDIES

1) EFFECTIVENESS OF EACH LEARNING SCHEME

This section observes performance depending on whether US-GAN's core components (i.e., feature decomposition and query selection-based contrastive learning) are applied or not. First, let's look at the qualitative performance depending on whether each component is applied or not. (c)-(e) in Fig. 7 is the result. (c) is the result without applying query selection-based contrastive loss $\mathcal{L}_{qs}$. Since the goal of $\mathcal{L}_{qs}$ is to boost the content preservation effect of the input image, it is observed that the organ structure inside the blue circle of (c) is not preserved well. However, because texture-focused translation was possible through feature decomposition, the style of the B domain is well expressed in the crop image. On the other hand, in the case of (d) where feature decomposition is not applied, many cells with low lateral resolution are observed in the crop image. From (c) and (d), we can qualitatively examine that feature decomposition and application of $\mathcal{L}_{qs}$ operate so that they meet our purpose. Similar trend is also observed in (e).

The preceding qualitative analysis is linked to the results in Table 2. From (c) of Fig. 7, we have already qualitatively confirmed that $\mathcal{L}_{qs}$ can boost content preservation performance. This is quantitatively re-confirmed in the second row of Table 2. The SSIM of the second row where $\mathcal{L}_{qs}$ is applied shows a difference of up to 0.47 compared to the SSIM of the baseline. Additionally, from (d) in Fig. 7, we confirmed that the proposed feature decomposition greatly contributes to fine texture translation. FID and RMSC performances support this. FID of the third row where feature decomposition was applied is as low as 65 compared to the baseline. Additionally, it is also worth noting that RMSC is 6.8 higher than the baseline in terms of contrast, which is one of the textures in the B domain. Lastly, the performance of the last row where both are applied shows the best performance across all metrics.

2) ADDITIONAL VERIFICATION

In order to assess the generalization performance of US-GAN, we use an open-access dataset [33]. Fig. 8 illustrates the results of translating T1 MRI images into T2 MRI images. The primary difference between T1 and T2 is the inversion of brightness, which is well-preserved in the results by US-GAN while maintaining the original structure of the input images. Furthermore, considering the crucial role of contrast enhancement in tumor detection, our results are remarkable.

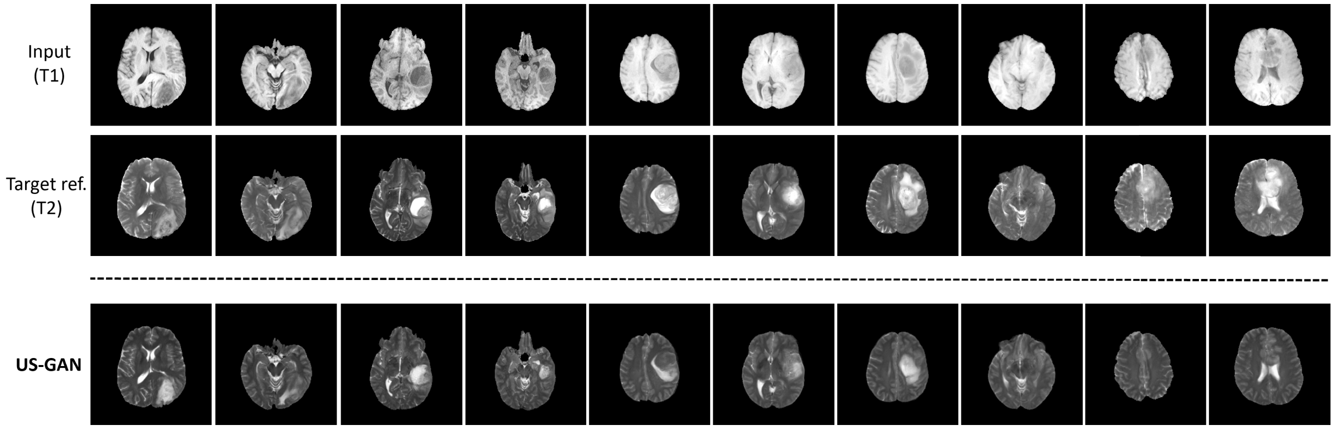


FIGURE 8. Additional qualitative results on MRI-dataset.

TABLE 3. Quantitative comparison on MRI dataset. Bold and underlined indicate the first and second place performances, respectively.

Methods	FID (↓)	SSIM (↑)	RMSC (↑)
CycleGAN [10]	74	0.62	36.1
CUT [11]	54	0.65	41.4
FSeSim [21]	51	0.66	37.4
Qs-attn [16]	43	0.79	40.0
US-GAN (ours)	**36**	**0.83**	**50.1**

We substantiate the generalization performance of qualitative results through a quantitative comparison using various metrics. Tab. 3 presents the quantitative results for BraTS. The FID score is 38 lower compared to CycleGAN. SSIM and RMSC values are respectively 0.21 and 14.0 higher than the lowest rank. Notably, the RMSC result indicates a significant improvement over the others in ultrasound images. These findings signify that the proposed method shows superior generalization not only in representation ability for the target domain but also in terms of content preservation compared to the baseline methods. Ultimately, it can be interpreted that the proposed feature decomposition and contrastive learning can yield excellent results across various transformations in medical image domain.

VI. CONCLUSION AND OTHER REMARKS

A. CONCLUSION

To successfully realize content preservation and fine texture translation, which are very important factors in ultrasound image-to-image translation (shortly, US I2I), we propose a novel US I2I framework that separates texture features and content features in latent space. Specifically, positive and negative example(s) are generated through multiple augmentations that have a visible effect on the texture, and a decomposition module trained with a famous contrastive objective successfully separates texture and content features. The excellence of the proposed method was experimentally verified through quantitative and qualitative performance analysis. The proposed method will be a great inspiration for future research in US I2I field in that it has succeeded in expressing fine cells while maintaining complex organ structures.

B. POTENTIAL SOCIETAL IMPACTS

The fact that US I2I can acquire images of various styles through one device will be of great help to the development of the medical field. For example, to diagnose a specific disease, style *B* may be a better option than *A*. However, despite this benefit, indiscriminate acquisition and modification of US images is fatal to patients' privacy. Therefore, in order for US I2I technology to be studied stably, the ethical awareness of engineers and researchers must be fostered.

REFERENCES

[1] X. Liu, J. Song, S. Wang, J. Zhao, and Y. Chen, "Learning to diagnose cirrhosis with liver capsule guided ultrasound image classification," *Sensors*, vol. 17, no. 12, p. 149, Jan. 2017.

[2] S. Sudharson and P. Kokil, "An ensemble of deep neural networks for kidney ultrasound image classification," *Comput. Methods Programs Biomed.*, vol. 197, Dec. 2020, Art. no. 105709.

[3] H. M. Balaha, M. Saif, A. Tamer, and E. H. Abdelhay, "Hybrid deep learning and genetic algorithms approach (HMB-DLGAHA) for the early ultrasound diagnoses of breast cancer," *Neural Comput. Appl.*, vol. 34, no. 11, pp. 8671–8695, Jun. 2022.

[4] Q. Huang, Y. Luo, and Q. Zhang, "Breast ultrasound image segmentation: A survey," *Int. J. Comput. Assist. Radiol. Surgery*, vol. 12, no. 3, pp. 493–507, Mar. 2017.

[5] Y. Xu, Y. Wang, J. Yuan, Q. Cheng, X. Wang, and P. L. Carson, "Medical breast ultrasound image segmentation by machine learning," *Ultrasonics*, vol. 91, pp. 1–9, Jan. 2019.

[6] Z. Jin, X. Li, Y. Zhang, L. Shen, Z. Lai, and H. Kong, "Boundary regression-based reep neural network for thyroid nodule segmentation in ultrasound images," *Neural Comput. Appl.*, vol. 34, no. 24, pp. 22357–22366, Dec. 2022.

[7] L. Teng, Z. Fu, and Y. Yao, "Interactive translation in echocardiography training system with enhanced cycle-GAN," *IEEE Access*, vol. 8, pp. 106147–106156, 2020.

[8] D. Tomar, L. Zhang, T. Portenier, and O. Goksel, "Content-preserving unpaired translation from simulated to realistic ultrasound images," in *Medical Image Computing and Computer Assisted Intervention–MICCAI*. Cham, Switzerland: Springer, 2021, pp. 659–669.

[9] H. Liu, J. Liu, S. Hou, T. Tao, and J. Han, "Perception consistency ultrasound image super-resolution via self-supervised CycleGAN," *Neural Comput. Appl.*, vol. 35, no. 17, pp. 12331–12341, Jun. 2023.

[10] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2242–2251.

[11] T. Park, A. A. Efros, R. Zhang, and J.-Y. Zhu, "Contrastive learning for unpaired image-to-image translation," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 319–345.

[12] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2414–2423.

[13] X. Chen, Y. Duan, R. Houthooft, J. Schulman, I. Sutskever, and P. Abbeel, "Infogan: Interpretable representation learning by information maximizing generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016, pp. 1–18.

[14] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, "Beta-vae: Learning basic visual concepts with a constrained variational framework," in *Proc. Int. Conf. Learn. Represent.*, 2016, pp. 1–17.

[15] H.-Y. Lee, H.-Y. Tseng, J.-B. Huang, M. Singh, and M.-H. Yang, "Diverse image-to-image translation via disentangled representations," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 35–51.

[16] X. Hu, X. Zhou, Q. Huang, Z. Shi, L. Sun, and Q. Li, "QS-Attn: Query-selected attention for contrastive learning in I2I translation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 18270–18279.

[17] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5967–5976.

[18] T.-C. Wang, M.-Y. Liu, J.-Y. Zhu, A. Tao, J. Kautz, and B. Catanzaro, "High-resolution image synthesis and semantic manipulation with conditional GANs," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8798–8807.

[19] M. Woo Kim and N. Ik Cho, "WHFL: Wavelet-domain high frequency loss for sketch-to-image translation," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2023, pp. 744–754.

[20] J.-Y. Zhu, R. Zhang, D. Pathak, T. Darrell, A. A. Efros, O. Wang, and E. Shechtman, "Toward multimodal image-to-image translation," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–19.

[21] C. Zheng, T.-J. Cham, and J. Cai, "The spatially-correlative loss for various image translation tasks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16402–16412.

[22] Y. Zhao, X. Shen, Z. Jin, H. Lu, and X.-S. Hua, "Attribute-driven feature disentangling and temporal aggregation for video person re-identification," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 4908–4917.

[23] T. H. Nguyen-Phuoc, C. Richardt, L. Mai, Y. Yang, and N. Mitra, "Block-gan: Learning 3D object-aware scene representations from unlabelled images," *Adv. neural Inf. Process. Syst.*, vol. 33, pp. 6767–6778, 2020.

[24] N. Behrmann, M. Fayyaz, J. Gall, and M. Noroozi, "Long short view feature decomposition via contrastive video representation learning," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9224–9233.

[25] S. Tulyakov, M.-Y. Liu, X. Yang, and J. Kautz, "MoCoGAN: Decomposing motion and content for video generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1526–1535.

[26] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2018, *arXiv:1807.03748*.

[27] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *Proc. 37th Int. Conf. Mach. Learn.*, vol. 119, Jul. 2020, pp. 1597–1607.

[28] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. A. Pires, Z. Guo, M. G. Azar, and B. Piot, "Bootstrap your own latent-a new approach to self-supervised learning," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, 2020, pp. 21271–21284.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–19.

[30] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul. 1948.

[31] X. Mao, Q. Li, H. Xie, R. Y. K. Lau, Z. Wang, and S. P. Smolley, "Least squares generative adversarial networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2813–2821.

[32] U. Demir and G. Unal, "Patch-based image inpainting with generative adversarial networks," 2018, *arXiv:1803.07422*.

[33] B. H. Menze et al., "The multimodal brain tumor image segmentation benchmark (BRATS)," *IEEE Trans. Med. Imag.*, vol. 34, no. 10, pp. 1993–2024, Oct. 2015.

[34] A. Paszke et al., "Pytorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[35] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–11.

[36] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–13.

[37] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2818–2826.

[38] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 248–255.

[39] E. Peli, "Contrast in complex images," *J. Opt. Soc. Amer. A, Opt. Image Sci.*, vol. 7, no. 10, p. 2032, 1990.

SEONGHO KIM (Associate Member, IEEE) received the B.S. degree in electronic engineering from Inha University, Incheon, South Korea, in 2022, where he is currently pursuing the M.S. degree in electrical and computer engineering. His research interests include image-to-image translation, facial expression manipulation, and video generation.

BYUNG CHEOL SONG (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees in electrical engineering from Korea Advanced Institute of Science and Technology (KAIST), Daejeon, South Korea, in 1994, 1996, and 2001, respectively. From 2001 to 2008, he was a Senior Engineer with Samsung Research (formerly, Digital Media Research and Development Center), Samsung Electronics Company Ltd., Suwon, South Korea. In 2008, he joined with the Department of Electronic Engineering, Inha University, Incheon, South Korea. He is currently a Professor. His research interests include the general areas of image processing and computer vision.