FEDERATED CONTRASTIVE LEARNING FOR PRIVACY-PRESERVING UNPAIRED IMAGE-TO-IMAGE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

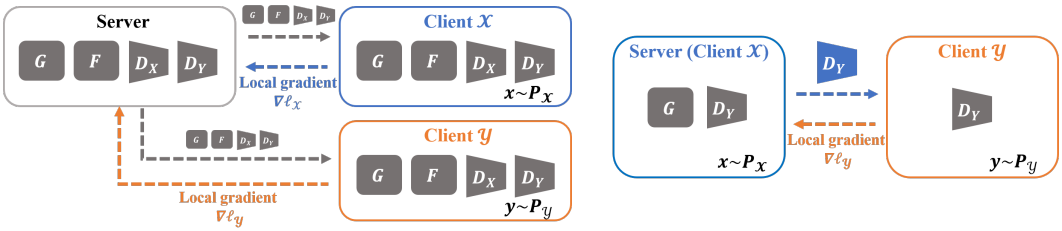
The goal of an unsupervised image-to-image translation (I2I) is to convert an input image in a specific domain to a target domain using a neural network trained with unpaired data. Existing I2I methods usually require a centrally stored dataset, which can compromise data privacy. A recent proposal of federated cycleGAN (FedCycleGAN) can protect the data-privacy by splitting the loss between the server and the clients so that the data does not need to be shared, but the weights and gradients of both generator and discriminators should be exchanged, demanding significant communication cost. To address this, here we propose a novel federated contrastive unpaired translation (FedCUT) approach for privacy-preserving image-to-image translation. Similar to FedCycleGAN, our method is based on the observation that the CUT loss can be decomposed into domain-specific local objectives, but in contrast to FedCycleGAN, our method only exchanges weights and gradients of a discriminator, significantly reducing the band-width requirement. In addition, by combining it with the pre-trained VGG network, the learnable part of the discriminator can be further reduced without impairing the image quality, resulting in two order magnitude reduction in the communication cost. Through extensive experiments for various translation tasks, we confirm that our method shows competitive performance compared to existing approaches.

1 INTRODUCTION

Unsupervised image-to-image (I2I) translation is to learn image conversion from one domain to another without matched training data from the two domains. Previous works (Zhu et al., 2017; Liu et al., 2017; Lee et al., 2018; Huang et al., 2018; Kim et al., 2020; Park et al., 2020; Tang et al., 2021) have shown great success in generating synthetic images which are realistic, and have been extended to multi-modal image-to-image translation (Huang et al., 2018), multi-domain image-to-image translation (Choi et al., 2018; Wu et al., 2019; Choi et al., 2020), and few-shot learning (Liu et al., 2019). However, most of the existing I2I translation methods are based on generator and discriminator architecture, which needs an access to a centralizing dataset that contains both input domain data and target domain data. This training scheme can be sometime against data privacy.

Recently, federated learning (FL) (McMahan et al., 2017) was proposed to protect data privacy by exchanging parameters of models between a server and clients without transmitting privately sensitive data. Specifically, the process of federated learning can be described in three steps. First, the server sends parameters of the global model to clients. Each client then trains its local model using personal data and sends its local update to the central server. Lastly, the server updates the parameters of the global model using aggregated information from multiple clients. The most representative algorithm is FedAvg (McMahan et al., 2017), which has inspired successful follow-up studies (Smith et al., 2017; Geyer et al., 2017; Zhao et al., 2018; Li et al., 2020; Wang et al., 2020).

Although most previous studies have been designed for classification problems or language modeling tasks in federated settings, several recent works have shown the possibility of using federated learning in generative models (Augenstein et al., 2020; Chen et al., 2020) and domain adaptation (Peng et al., 2020; Yao et al., 2021). However, there is still a lack of research on federated learning for unsupervised image-to-image translation, although federated version of unsupervised I2I is quite



(a) Federated CycleGAN (FedCycleGAN) (b) Federated CUT (FedCUT)

Figure 1: Comparison of FedCycleGAN and FedCUT.

useful in protecting copyright (Song & Ye, 2021) or medical imaging applications. For example, in multi-center x-ray computed tomography (CT) imaging studies, the scanner type, radiation dose, and filter kernels vary depending on each medical institutes, so the image translation to a normalized one is necessary for robust statistical analysis (Vegas-Sánchez-Ferrero et al., 2019; Selim et al., 2021). In this case, federated I2I is desirable as each participating hospital does not need to share their data, but still get the normalized ones.

Recently, federated CycleGAN (FedCycleGAN) (Song & Ye, 2021) was proposed to solve unsupervised image-to-image translation in federated learning environment. FedCycleGAN does not require any private data exchange based on the fact that CycleGAN (Zhu et al., 2017) loss can be broken down to domain-specific local objectives. Specifically, the server in FedCycleGAN transmits the parameters of global models, while each client sends its local gradient calculated with its local data. The server then updates the parameters of global networks using local gradients from clients. Unfortunately, FedCycleGAN training requires transmission of both generators and discriminators, which can be a bottleneck in federated learning environment (see Fig. 1(a)).

To overcome this limitation, here we propose a novel federated contrastive unpaired translation (FedCUT). The key idea of FedCUT is the decomposition of recent contrastive unpaired translation (CUT) (Park et al., 2020) loss into domain-specific objectives. Specifically, the total loss of CUT is the sum of the local objective for the input domain and another local objective for the target domain. Accordingly, client can calculate a domain-specific local objective using its data and transmit gradient information to the server without sharing personal data. In particular, as shown in Fig. 1(b), FedCUT only exchanges parameters and local gradients of the discriminator D_Y, which require a lower bandwidth than FedCycleGAN. Furthermore, the discriminator architecture can be further simplified based on a pre-trained classification network such as VGGNet (Simonyan & Zisserman, 2015), resulting in two order of magnitude smaller bandwidth requirement compared to the FedCycleGAN.

Despite the simplification, experimental results show that our method achieves comparable results compared to existing baselines in various unsupervised image-to-image translation tasks. The main contribution of this paper is as follows:

1. Based on a novel observation that CUT loss can be decomposed to domain-specific objectives, we propose FedCUT which shows better performance and lower bandwidth requirement than those of conventional federated image-to-image translation method.
2. By using a simplified discriminator based on a pre-trained VGG classification network, we achieve competitive image generation performance in spite of reducing the transmission overhead by two order of magnitude compared to the existing approach.

2 Related Work

2.1 Federated CycleGAN

The goal of CycleGAN is to learn to translate a image from one domain ($\mathcal{X}$) to a corresponding output image in another domain ($\mathcal{Y}$). Suppose that $P_{\mathcal{X}}$ is a probability distribution of $\mathcal{X}$ and $P_{\mathcal{Y}}$ is that of $\mathcal{Y}$ and x, and y are images from $\mathcal{X}$ and $\mathcal{Y}$, respectively. The generator $G : \mathcal{X} \mapsto \mathcal{Y}$ translates an image from $\mathcal{X}$ to an output image in $\mathcal{Y}$. The discriminator D_Y distinguishes real samples in $\mathcal{Y}$

and fake samples that are generated by G using samples in $\mathcal{X}$. Similarly, $F : \mathcal{Y} \mapsto \mathcal{X}$ is the generator that translates an image in $\mathcal{Y}$ into a corresponding output in $\mathcal{X}$. The discriminator D_X distinguishes real images in $\mathcal{X}$ from fake images that are made by F using images in $\mathcal{Y}$.

In the CycleGAN, the following minmax optimization problem needs be solved:

$$\min_{G,F} \max_{D_X,D_Y} \ell_{CycleGAN}(G, F, D_X, D_Y). \tag{1}$$

Here, the total loss is composed of adversarial loss and the cycle-consistency loss:

$$\ell_{CycleGAN}(G, F, D_X, D_Y) = \ell_{GAN}(G, D_Y) + \ell_{GAN}(F, D_X) + \lambda \ell_{cycle}(G, F) \tag{2}$$

where the adversarial losses are given by

$$\begin{aligned} \ell_{GAN}(G, D_Y) &= \mathbb{E}_{y\sim P_{\mathcal{Y}}}[\log D_Y(y)] + \mathbb{E}_{x\sim P_{\mathcal{X}}}[\log(1 - D_Y(G(x)))] \\ \ell_{GAN}(F, D_X) &= \mathbb{E}_{x\sim P_{\mathcal{X}}}[\log D_X(x)] + \mathbb{E}_{y\sim P_{\mathcal{Y}}}[\log(1 - D_X(F(y)))] \end{aligned} \tag{3}$$

and the cycle-consistency loss is

$$\ell_{cycle}(G, F) = \mathbb{E}_{x\sim P_{\mathcal{X}}}[||F(G(x)) - x||_1] + \mathbb{E}_{y\sim P_{\mathcal{Y}}}[||G(F(y)) - y||_1], \tag{4}$$

The key idea of the FedCycleGAN (Song & Ye, 2021) is the following loss decomposition:

$$\ell_{CycleGAN}(G, F, D_X, D_Y) = \ell_{\mathcal{X}}(G, F, D_X, D_Y) + \ell_{\mathcal{Y}}(G, F, D_X, D_Y) \tag{5}$$

where $\ell_{\mathcal{X}}$ and $\ell_{\mathcal{Y}}$ are local objectives that only use data in $\mathcal{X}$ and $\mathcal{Y}$ domains, respectively:

$$\begin{aligned} \ell_{\mathcal{X}}(G, F, D_X, D_Y) &= \mathbb{E}_{x\sim P_{\mathcal{X}}}[\log D_X(x)] + \mathbb{E}_{x\sim P_{\mathcal{X}}}[\log(1 - D_Y(G(x)))] \\ &\quad + \lambda \mathbb{E}_{x\sim P_{\mathcal{X}}}[||F(G(x)) - x||_1] \end{aligned} \tag{6}$$

$$\begin{aligned} \ell_{\mathcal{Y}}(G, F, D_X, D_Y) &= \mathbb{E}_{y\sim P_{\mathcal{Y}}}[\log(1 - D_X(F(y)))] + \mathbb{E}_{y\sim P_{\mathcal{Y}}}[\log D_Y(y)] \\ &\quad + \lambda \mathbb{E}_{y\sim P_{\mathcal{Y}}}[||G(F(y)) - y||_1] \end{aligned} \tag{7}$$

Accordingly, the client with the data in $\mathcal{X}$ domain uses $\ell_{\mathcal{X}}$ with its own data, whereas the other client in $\mathcal{Y}$ domain employs the loss $\ell_{\mathcal{Y}}$ without revealing its data. Then, the server can train the global models by using the local gradients without accessing the data itself. However, as shown in Eqs. 6 and 7, each client need knowledge of both generators and discriminators, resulting in high communication costs.

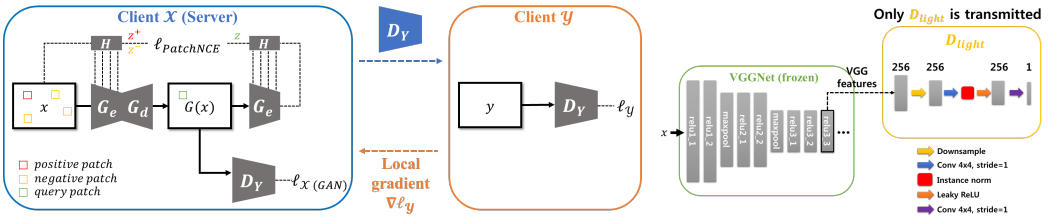
3 FEDERATED CUT

3.1 CONTRASTIVE UNPAIRED TRANSLATION

As a simple but a power alternative to cycleGAN, contrastive unpaired translation (CUT) (Park et al., 2020) was recently proposed. As shown in Fig. 2(a) and the detailed explanation in the image caption, CUT is composed of the generator $G : \mathcal{X} \mapsto \mathcal{Y}$ and the discriminator D_Y. The generator G consists of the encoder G_e and the decoder G_d and translates the domain of input images $\mathcal{X}$ to the target domain $\mathcal{Y}$. The discriminator D_Y discriminates a synthetic images $G(x)$ from a real image y in the target domain $\mathcal{Y}$.

In contrast to cycleGAN, CUT does not require cycle-consistency. Instead, a contrastive loss is used to impose one-to-one correspondency. Specifically, a multi-layer and patch-wise application of InfoNCE loss (Oord et al., 2018), dubbed as PatchNCE loss, is employed as a contrastive loss. The purpose of PatchNCE loss is to match an input patch (positive) and the output patch (query) at a specific location. Non-corresponding patches are regarded as negatives. To extract meaningful vectors mapped from query, positive, and negatives, the encoder G_e and a two-layer MLP network H can be used. As shown in Fig. 2(a), we denotes ${z_l}^{s+} = H_l({G_e}^l(x))$ by a feature vector from an input, where ${G_e}^l$ is the l-th selected layers from the encoder G_e, H_l is the corresponding MLP, and s denotes a specific location. Similarly, ${z_l}^{s-}$ is a feature vector from a negative, and ${z_l}^s = H_l({G_e}^l(G(x)))$ is a feature vector from an output. Using these feature vectors, PatchNCE loss $\ell_{PatchNCE}$ can be calculated as follows:

$$\ell_{PatchNCE}(G, H, X) = \mathbb{E}_{x\sim P_{\mathcal{X}}} \left[\sum_{l=1}^{L} \sum_{s=1}^{S_l} \ell_{NCE}(z_l{}^s, z_l{}^{s+}, z_l{}^{s-}) \right]. \tag{8}$$



(a) Federated CUT (FedCUT) (b) Architecture of VGGNet-based discriminator

Figure 2: Architectures of (a) FedCUT and (b) VGGNet-based discriminator. In (a), if the y image is inserted as additional input to D_Y at the client $\mathcal{X}$, the architecture in client $\mathcal{X}$ is equivalent to the original CUT (Park et al., 2020).

where L is the last index for the selection of layers of the encoder G_e, and S_l is the number of locations for the l-th selected layers. Furthermore, in order to avoid unnecessary changes, CUT use $\ell_{PatchNCE}(G, H, Y)$ using images from target domain $\mathcal{Y}$, which can be seen as the identity loss for the target domain $\mathcal{Y}$. The total objective for CUT is as follows:

$$\begin{aligned}\ell_{CUT}(G, H, D_Y, X, Y) = \ell_{GAN}(G, D_Y, X, Y) + \lambda_X \ell_{PatchNCE}(G, H, X) \\ + \lambda_Y \ell_{PatchNCE}(G, H, Y).\end{aligned} \tag{9}$$

3.2 Derivation of FedCUT

In FastCUT, a variant of CUT, $\lambda_Y = 0$ is used without imposing the identity loss (Park et al., 2020). In this case, the total loss for CUT can be decomposed into domain-specific local objectives:

$$\ell_{CUT}(G, H, D_Y, X, Y) = \ell_{\mathcal{X}}(G, H, D_Y, X) + \ell_{\mathcal{Y}}(D_Y, Y) \tag{10}$$

where $\ell_{\mathcal{X}}$ is the $\mathcal{X}$ domain-specific local objective and $\ell_{\mathcal{Y}}$ is the local loss for the target domain $\mathcal{Y}$:

$$\ell_{\mathcal{X}}(G, H, D_Y, X) = \mathbb{E}_{x\sim P_{\mathcal{X}}}[\log(1 - D_Y(G(x)))] + \lambda_X \ell_{PatchNCE}(G, H, X) \tag{11}$$

$$\ell_{\mathcal{Y}}(D_Y, Y) = \mathbb{E}_{y\sim P_{\mathcal{Y}}}[\log D_Y(y)] \tag{12}$$

Accordingly, we can develop a federated CUT (FedCUT), where each client calculates its domain-specific loss using its own data without data exchange as shown in Fig. 2(a). Note that only the discriminator is shared in Eqs. 11 and 12. Accordingly, the training process for FedCUT is as follows: the server (the client whose domain is $\mathcal{X}$) sends the parameters of the discriminator D_Y to another client which has data from the target domain $\mathcal{Y}$. Then the client (the client whose domain is $\mathcal{Y}$) calculates the domain-specific loss $\ell_{\mathcal{Y}}$ and sends gradient information of the discriminator D_Y to the server. Next, the server updates the generator G and the discriminator D_Y using its domain-specific loss $\ell_{\mathcal{X}}$ and gradients from the client. This process is repeated until the convergence of networks. Compared to conventional FedCycleGAN that needs to transfer both two generators and two discriminators, the transmission of the discriminator parameters and gradients is necessary.

3.3 Discriminator simplification

If the number of parameters of the discriminator D_Y can be reduced, communication costs can be also reduced, but a high capacity of the discriminator D_Y is also critical for successful training of the generator G. Accordingly, the standard FedCUT uses the PatchGAN (Isola et al., 2017) structure for the discriminator D_Y.

In this paper, inspired by (Sungatullina et al., 2018), we construct a discriminator based on a pre-trained classification network. Specifically, we construct the discriminator by using the pre-trained VGGNet (Simonyan & Zisserman, 2015) as the backbone and adding the lightweight network D_{light} consisting of several convolutional layers. The advantage of using a pre-trained classification network is that rich perceptual features can be used to train a discriminator, resulting in a better representation power than a randomly initialized discriminator (Sungatullina et al., 2018). As the pre-trained VGGNet is known to clients and the server, we only need to transmit D_{light} for federated learning, resulting in low communication costs.

Fig. 2(b) shows the architecture of the VGGNet-based discriminator. Pre-trained VGGNet extracts perceptual features, which are followed by a downsampling layer and several convolutional layers. For downsampling, we use a 3×3 Gaussian filter to prevent the generation of high frequency artifacts. We use pre-trained VGG16 using ImageNet (Deng et al., 2009), and the parameters of VGGNet are fixed during a training process. We choose the '$relu3\_3$' layer of VGG16 to extract features that are followed by a lightweight discriminator D_{light}.

Table 1: Communication cost of federated learning.

	Networks to be transmitted	Communication cost (byte)
FedCycleGAN	G, F, D_X, D_Y	1.80×10^8
FedCUT	D_{light}	4.20×10^6

Table 1 shows communication costs of FedCycleGAN, and FedCUT for in various image-to-image translations tasks. Communication cost in Table 1 represents the cost for sending parameters of networks between a server and a client. Our FedCUT significantly reduces the bandwidth requirement by two orders of magnitude compared to FedCycleGAN.

4 METHODS

4.1 DATASET

Natural image translation To evaluate the performance of our methods, we first conducted various image-to-image translation tasks using three natural image datasets: horse-to-zebra, cat-to-dog, and cityscapes dataset (Cordts et al., 2016). Horse-to-zebra dataset was used in the original CycleGAN (Zhu et al., 2017) and consists of training images (1067 horse and 1334 zebra images) and test data (120 horse and 140 zebra images) from ImageNet (Deng et al., 2009). Cat-to-dog dataset is divided into training data (5153 cat and 4739 dog images) and 500 test images from AFHQ dataset (Choi et al., 2020). Cityscapes dataset contains cityscapes and corresponding segmentation maps and consists of 2975 training data and 500 test data. As in the experimental setting of CUT (Park et al., 2020), we use images with a resolution of 256×256 for the training and the test. We use unpaired images for unsupervised image-to-image translation in all experiments.

CT image translation X-ray computed tomography (CT) is an important imaging system for radiological diagnosis. Since a high dose of radiation carries a risk of cancer, the dose reduction is quite often used. Unfortunately, a high level of noise in the low-dose CT scan can lead to misdiagnosis so that low-dose CT noise reduction is an important research topic in the field of medical imaging. Recently, unpaired image translation using cycleGAN was demonstrated effective for low-dose noise image reduction by converting the denoising problem as an image translation from low-dose to high-dose CT images (Kang et al., 2019). In our experiment, we assume that one hospital (server) only has low-dose CT images while another hospital (client) has routine-dose CT scans, and use the proposed FedCUT to conduct denoising experiment.

To conduct the low-dose CT denoising experiment, we utilize AAPM CT dataset as in previous works (Kang et al., 2017; 2018) made by using CT data from AAPM 2016 Low Dose CT Grand Challenge (McCollough et al., 2017). We use CT scans from 8 patients for training, while CT scans from one patient are used for test data. The training data consists 3236 slices and 350 CT slices are used for the test. The resolution of CT images is 512×512 pixels.

4.2 BASELINES

Natural image translation For comparison, we obtained various image-to-image translation results using non-federated methods and federated methods, such as (1) CUT (Park et al., 2020), (2) FastCUT (Park et al., 2020), (3) FedCycleGAN (Song & Ye, 2021), and (4) our FedCUT.

For a fair comparison, we follow the default setting of FastCUT (Park et al., 2020). Specifically, we use ResNet-based generator (He et al., 2016) and PatchGAN (Isola et al., 2017) as the discriminator with the LSGAN loss (Mao et al., 2017). The encoder is the half of the generator and we utilize features from five selected layers to calculate PatchNCE loss as in the setting of CUT (Park et al., 2020). Only FedCUT uses the different discriminator, as explained in Section 3.3. To train FedCUT,

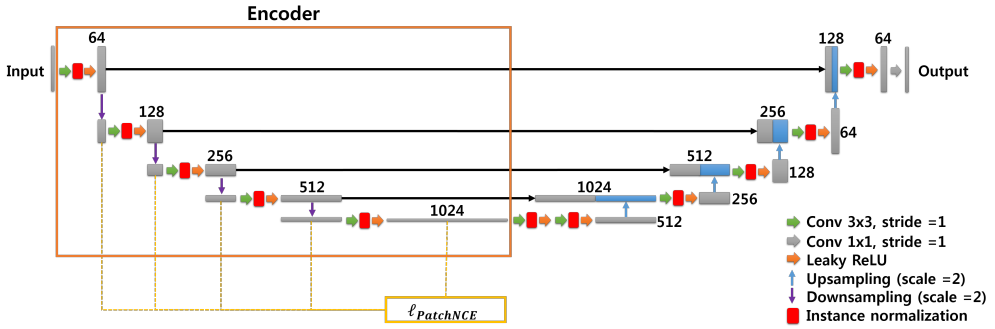


Figure 3: Architectures of the generator and feature extractor for PatchNCE loss in our low-dose CT denoising experiments.

Table 2: Quantitative comparison with various methods for natural image translation tasks.

	Horse-to-Zebra	Cat-to-Dog	Cityscapes			
	FID ↓	FID ↓	FID ↓	mAP ↑	pixAcc ↑	classAcc ↑
CUT	**45.5**	76.2	56.4	24.7	68.8	30.7
FastCUT	73.4	94.0	68.8	19.1	59.9	24.3
FedCycleGAN	72.2	98.5	58.3	22.9	66.5	31.8
FedCUT	55.5	**75.2**	**54.8**	**26.5**	**72.9**	**34.5**

we use the Adam optimizer (Kingma & Ba, 2014) using $\beta_1 = 0.5$ and $\beta_2 = 0.999$ for 200 epochs. The learning rate is 0.002, which is fixed for the first 150 epochs and then gradually decreases to 0 for the remaining epochs. The batch size is 1. We also use same augmentation strategy of FastCUT (Park et al., 2020). In particular, we resize an image to 286×286 images and crop a 256×256 image, which are then flipped horizontally at random. FastCUT use $\lambda_X = 10$ and $\lambda_Y = 0$ in Eq. 9. The hyperparameter of FedCUT was set $\lambda_X = 1$ in Eq. 10 as they gave the optimal performance in our data set. We utilize Pytorch (Paszke et al., 2019) library and a NVIDIA GeForce RTX 3090 for the implementation.

CT image translation For performance evaluation in federated setting, for CT image translation tasks, we compare FedCycleGAN, and FedCUT. We use PatchGAN (Isola et al., 2017) as the discriminator in FedCycleGAN as in the original paper (Song & Ye, 2021), and the VGGNet-based discriminator in our FedCUT. For a generator, we use U-net (Ronneberger et al., 2015) generator which shows successful performance in medical image processing. Fig. 3 shows the architecture of the generator for low-dose CT denoising. We define the encoder as the half of the Unet, from which features are extracted to compute the PatchNCE loss as shown in as in Fig. 3. For data augmentation, we crop the patches with the size of 128×128 pixels from 512×512 images. In addition, we apply horizontal and vertical flip to patches. Other training setting is same as that of natural image translation experiments.

Note that FedCycleGAN (Song & Ye, 2021) use the identity loss (Zhu et al., 2017) to prevent artifacts outside the object in CT images. Instead, we use the background consistency module (BCM) loss (Du et al., 2020), which preserves the background part of images but does not require generator transmission.

4.3 Evaluation Metrics

Natural image translation We adopt the evaluation protocol of CUT (Paszke et al., 2019). To evaluate the performance of our methods, we calculate Frechet Inception Distance (FID) (Heusel et al., 2017). For Cityscapes dataset, we pre-train DRN (Yu et al., 2017), which produces semantic segmentation maps. We then use the pre-trained DRN to produce semantic segmentation maps from the generated images. Based on ground truths of segmentation maps, we calculate mean average precision (mAP), average class accuracy (classAcc), pixel-wise accuracy (pixAcc).

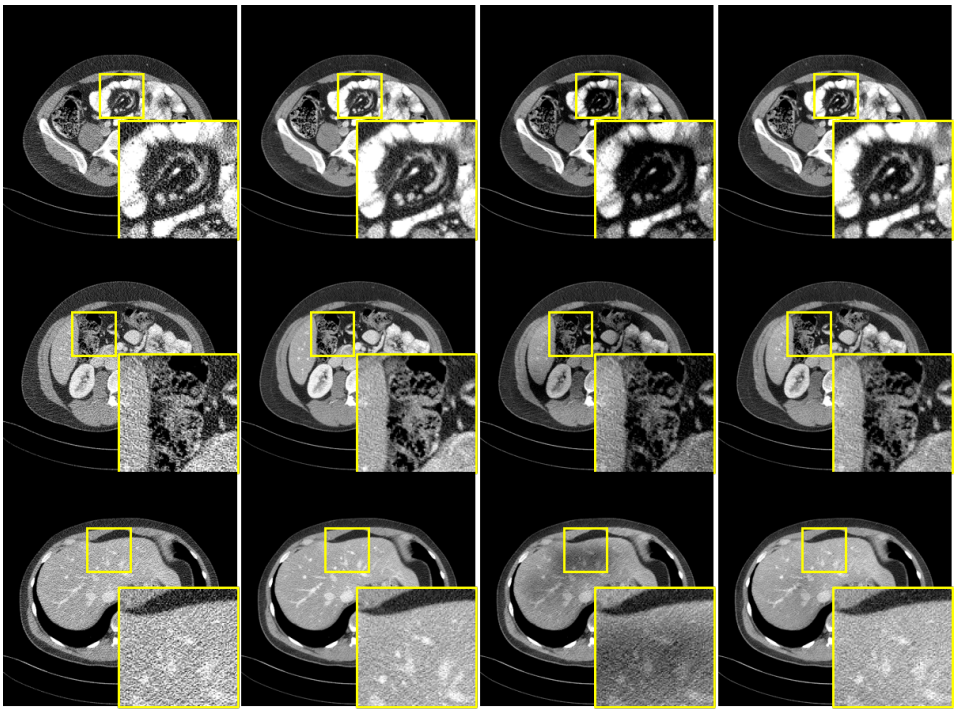
Figure 4: Image-to-image translation results for horse-to-zebra, cat-to-dog, and cityscapes.

CT image translation To evaluate the denoising performance of methods, we calculate the peak signal-to-noise ratio (PSNR) and the structural similarity index metric (SSIM) (Wang et al., 2004). In addition, we compare the communication costs of the individual methods.

5 Experimental Results

5.1 Qualitative and Quantitative Evaluation

Natural image translation Fig. 4 shows qualitative results for various image-to-image translation tasks. Surprisingly, our FedCUT shows visually pleasing results while other methods sometimes make unrealistic results. In particular, CUT creates artificial structures around the animal's mouth that are different from the sources, while FedCUT maintains the input contents and produces the appearance similar to the target domain. Overall FedCUT produces more realistic image details compared to other methods. These properties may be due to the use of rich perceptual features that give networks a better chance of learning desirable statistics. Table 2 shows quantitative results of various methods. FedCUT performs better than other methods on cat-to-dog and Cityscapes tasks



(a) Low-dose CT (b) Routine-dose CT (c) FedCycleGAN (d) FedCUT

Figure 5: Low-dose CT denoising results. Intensity range is (-160, 240) [HU] (Hounsfield Unit).

Table 3: Quantitative comparison for low-dose CT denoising.

	PSNR [dB] ↑	SSIM ↑	Communication cost (byte)
Input	32.5132	0.7411	·
FedCycleGAN	34.2794	**0.8218**	2.34×10^8
FedCUT	**35.7753**	0.8113	4.20×10^6

and has a lower FID than FastCUT and FedCycleGAN on horse-to-zebra task. This may be also thanks to the rich texture features from the pre-trained VGGNet.

CT image translation Fig. 5 shows qualitative results for low-dose CT denoising. While FedCycleGAN sacrifices fine structures of images, our method shows better denoising results preserving details. In addition, as shown in Table 3, quantitative results confirmed that our methods achieve higher PSNR values than FedCycleGAN. Although VGG is pre-trained with ImageNet, which has different domain data from CT images, FedCUT still shows the highest PSNR. This shows that features learned from natural images can also offer useful features for solving problems in medical imaging (Ciompi et al., 2015; Altaf et al., 2019). Although FedCycleGAN produces the best SSIM score, the communication cost is impractical in a federated learning environment. However, our FedCUT requires much less communication costs compared to FedCycleGAN, making it more suitable for use in a federated learning environment.

5.2 Ablation Study

Different discriminator architecture To investigate the change in performance when using different discriminator structures, we also performed image translation tasks with FedCUT using PatchGAN (FedCUT$_{PatchGAN}$). Note that PatchGAN is used in the original CUT (Park et al., 2020). Table 4 shows performance comparisons when using different discriminators. We find that our FedCUT performs better than FedCUT$_{PatchGAN}$ on horse-to-zebra, cat-to-dog, and CT denoising tasks with lower communication costs.

Table 4: Ablation study using different discriminators.

	Horse-to-Zebra	Cat-to-Dog	Cityscapes				CT denoising		Communication
	FID ↓	FID ↓	FID ↓	mAP ↑	pixAcc ↑	classAcc ↑	PSNR [dB] ↑	SSIM ↑	cost (byte)
FedCUT$_{PatchGAN}$	59.6	82.2	**50.5**	26.0	72.1	**36.1**	35.2415	0.8048	1.11×10^7
FedCUT	**55.5**	**75.2**	54.8	**26.5**	**72.9**	34.5	**35.7753**	**0.8113**	4.20×10^6

Different VGG features We also conduct image translation tasks with FedCUT using various features from different layers of VGG16. Table 5 shows results when using different VGG features such as '$relu3\_1$', '$relu3\_2$', and our default '$relu3\_3$'. Our FedCUT achieves the highest scores in natural image translation tasks. Although FedCUT$_{relu3_2}$ produces better PSNR and SSIM values than FedCUT$_{default}$ in the CT denoising task, which implies that different VGG features need to be selected for high performance depending on the application. our FedCUT using '$relu3\_3$' can be applied for general purposes, as it shows better performance than FedCycleGAN on all tasks in our experiment with lower communication cost.

Table 5: Ablation study using different VGG features.

	Horse-to-Zebra	Cat-to-Dog	Cityscapes				CT denoising	
	FID ↓	FID ↓	FID ↓	mAP ↑	pixAcc ↑	classAcc ↑	PSNR [dB] ↑	SSIM ↑
FedCUT$_{relu3_1}$	60.7	93.9	55.4	24.0	70.6	33.2	35.7632	0.8105
FedCUT$_{relu3_2}$	63.9	84.7	82.0	22.1	69.3	30.6	**36.0510**	**0.8256**
FedCUT	**55.5**	**75.2**	**54.8**	**26.5**	**72.9**	**34.5**	35.7753	0.8113

6 CONCLUSIONS

In this paper, we propose a federated contrastive unpaired translation for privacy-preserving image-to-image translation (FedCUT). Thanks to the decomposition of the CUT loss into domain-specific local objectives, our framework can be used for unsupervised image-to-image translation without sharing data between a server and clients. Furthermore, FedCUT only requires the transmission of parameters of a discriminator, which requires low communication costs compared to the previous method. In addition, by using simplified discriminator based on a pre-trained VGGNet, FedCUT reduces communication costs by two order of magnitude while increasing the performance. Our experiments show that our method produces results that are comparable or even outperform the existing methods in several tasks despite the low communication costs.

REFERENCES

Fouzia Altaf, Syed MS Islam, Naveed Akhtar, and Naeem Khalid Janjua. Going deep in medical image analysis: concepts, methods, challenges, and future directions. *IEEE Access*, 7:99540–99572, 2019.

Sean Augenstein, H. Brendan McMahan, Daniel Ramage, Swaroop Ramaswamy, Peter Kairouz, Mingqing Chen, Rajiv Mathews, and Blaise Aguera y Arcas. Generative models for effective ml on private, decentralized datasets. In *International Conference on Learning Representations*, 2020.

Dingfan Chen, Tribhuvanesh Orekondy, and Mario Fritz. Gs-wgan: A gradient-sanitized approach for learning differentially private generators. In *Neural Information Processing Systems (NeurIPS)*, 2020.

Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

Francesco Ciompi, Bartjan de Hoop, Sarah J van Riel, Kaman Chung, Ernst Th Scholten, Matthijs Oudkerk, Pim A de Jong, Mathias Prokop, and Bram van Ginneken. Automatic classification of pulmonary peri-fissural nodules in computed tomography using an ensemble of 2d views and a convolutional neural network out-of-the-box. *Medical image analysis*, 26(1):195–202, 2015.

Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255, 2009.

Wenchao Du, Hu Chen, and Hongyu Yang. Learning invariant representation for unsupervised image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14483–14492, 2020.

Robin C Geyer, Tassilo Klein, and Moin Nabi. Differentially private federated learning: A client level perspective. *arXiv preprint arXiv:1712.07557*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *ECCV*, 2018.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

Eunhee Kang, Junhong Min, and Jong Chul Ye. A deep convolutional neural network using directional wavelets for low-dose x-ray CT reconstruction. *Medical physics*, 44(10):e360–e375, 2017.

Eunhee Kang, Won Chang, Jaejun Yoo, and Jong Chul Ye. Deep convolutional framelet denosing for low-dose CT via wavelet residual network. *IEEE transactions on medical imaging*, 37(6): 1358–1369, 2018.

Eunhee Kang, Hyun Jung Koo, Dong Hyun Yang, Joon Bum Seo, and Jong Chul Ye. Cycle-consistent adversarial denoising network for multiphase coronary ct angiography. *Medical physics*, 46(2):550–562, 2019.

Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwang Hee Lee. U-gat-it: Unsupervised generative attentional networks with adaptive layer-instance normalization for image-to-image translation. In *International Conference on Learning Representations*, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *European Conference on Computer Vision*, 2018.

Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. In I. Dhillon, D. Papailiopoulos, and V. Sze (eds.), *Proceedings of Machine Learning and Systems*, volume 2, pp. 429–450, 2020.

Ming-Yu Liu, Thomas Breuel, and Jan Kautz. Unsupervised image-to-image translation networks. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Ming-Yu Liu, Xun Huang, Arun Mallya, Tero Karras, Timo Aila, Jaakko Lehtinen, and Jan Kautz. Few-shot unsueprvised image-to-image translation. In *arxiv*, 2019.

Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2794–2802, 2017.

Cynthia H McCollough, Adam C Bartley, Rickey E Carter, Baiyu Chen, Tammy A Drees, Phillip Edwards, David R Holmes III, Alice E Huang, Farhana Khan, Shuai Leng, et al. Low-dose CT for the detection and classification of metastatic liver lesions: results of the 2016 low dose CT grand challenge. *Medical physics*, 44(10):e339–e352, 2017.

Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Aarti Singh and Jerry Zhu (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, volume 54 of *Proceedings of Machine Learning Research*, pp. 1273–1282. PMLR, 20–22 Apr 2017.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Taesung Park, Alexei A. Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European Conference on Computer Vision*, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, pp. 8024–8035, 2019.

Xingchao Peng, Zijun Huang, Yizhe Zhu, and Kate Saenko. Federated adversarial domain adaptation. In *International Conference on Learning Representations*, 2020.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

Md Selim, Jie Zhang, Baowei Fei, Guo-Qiang Zhang, and Jin Chen. Ct image harmonization for enhancing radiomics studies. *arXiv preprint arXiv:2107.01337*, 2021.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

Virginia Smith, Chao-Kai Chiang, Maziar Sanjabi, and Ameet S Talwalkar. Federated multi-task learning. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Joonyoung Song and Jong Chul Ye. Federated cyclegan for privacy-preserving image-to-image translation. *arXiv preprint arXiv:2106.09246*, 2021.

Diana Sungatullina, Egor Zakharov, Dmitry Ulyanov, and Victor Lempitsky. Image manipulation with perceptual discriminators. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 579–595, 2018.

Hao Tang, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE Transactions on Neural Networks and Learning Systems*, 2021.

Gonzalo Vegas-Sánchez-Ferrero, Maria Jesus Ledesma-Carbayo, George R Washko, and Raul San Jose Estepar. Harmonization of chest ct scans for different doses and reconstruction methods. *Medical physics*, 46(7):3117–3132, 2019.

Hongyi Wang, Mikhail Yurochkin, Yuekai Sun, Dimitris Papailiopoulos, and Yasaman Khazaeni. Federated learning with matched averaging. In *International Conference on Learning Representations*, 2020.

Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Po-Wei Wu, Yu-Jing Lin, Che-Han Chang, Edward Y. Chang, and Shih-Wei Liao. Relgan: Multi-domain image-to-image translation via relative attributes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019.

Chun-Han Yao, Boqing Gong, Yin Cui, Hang Qi, Yukun Zhu, and Ming-Hsuan Yang. Federated multi-target domain adaptation. *arXiv preprint arXiv:2108.07792*, 2021.

Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 472–480, 2017.

Yue Zhao, Meng Li, Liangzhen Lai, Naveen Suda, Damon Civin, and Vikas Chandra. Federated learning with non-iid data. *arXiv preprint arXiv:1806.00582*, 2018.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision (ICCV)*, pp. 2223–2232, 2017.

A Appendix

A.1 Algorithm 1: FedCUT

Algorithm 1: FedCUT

Input : total number of rounds (epochs) K, learning rate η
Output: generator $G(\cdot;\theta_G)$, MLP network $H(\cdot;\theta_H)$, discriminator $D_Y(\cdot;\theta_{D_Y})$

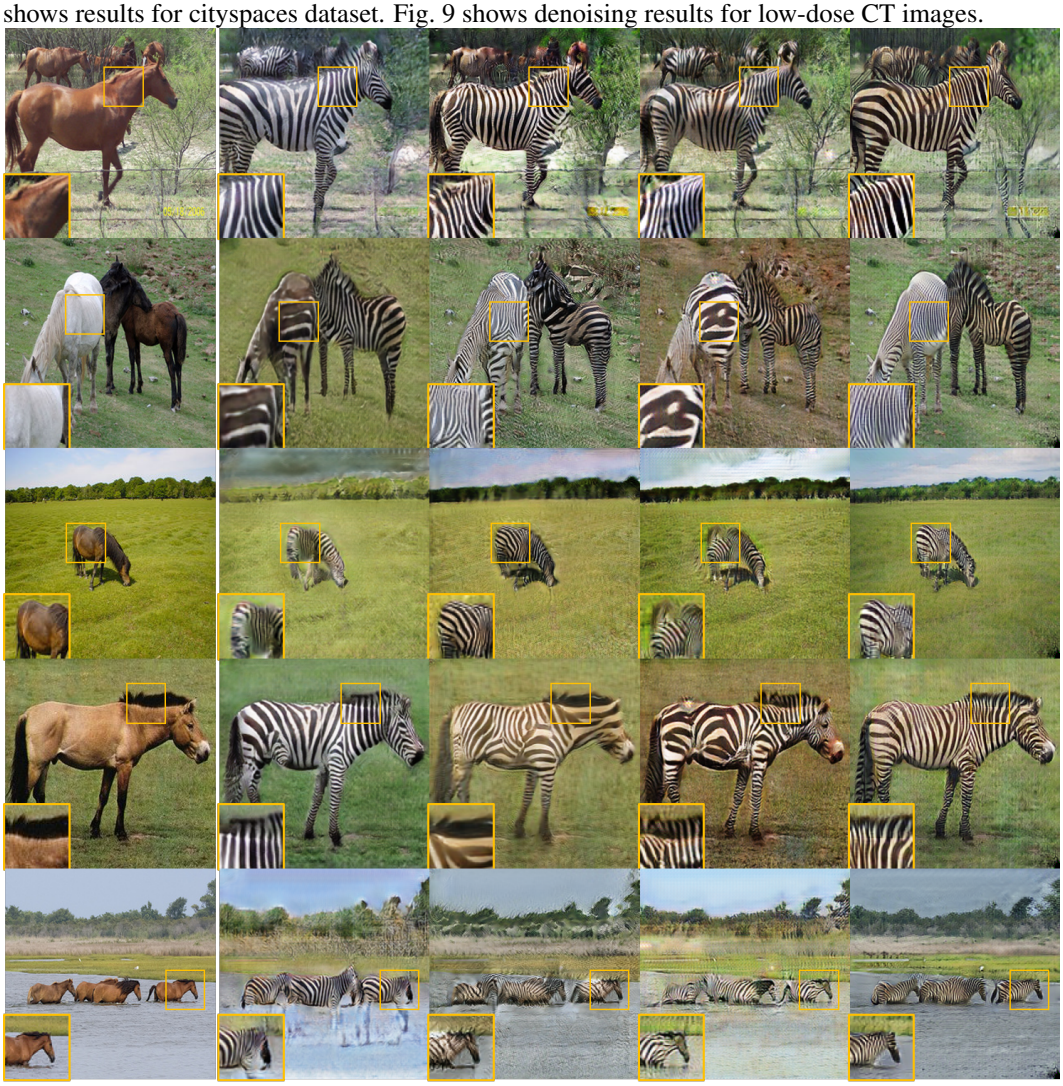
Initialize θ_G, θ_H, θ_{D_Y} in server $\mathcal{X}$;
for *each round k = 1, 2, ..., K* **do**
  $g_{\theta_{D_{Y,real}}} \leftarrow$ `GetLocalGradientFromClient` $(\theta_{D_Y}^{(k-1)})$;
  $b_X \leftarrow$ (sample batch from server $\mathcal{X}$) ;
  $g_{\theta_{D_{Y,fake}}} \leftarrow \nabla_{\theta_{D_Y}} \ell_{\mathcal{X}}(G, D_Y, b_X)$;
  ${\theta_{D_Y}}^{(k)} \leftarrow {\theta_{D_Y}}^{(k-1)} - \eta(g_{\theta_{D_{Y,real}}} + g_{\theta_{D_{Y,fake}}})$;
  ${\theta_G}^{(k)} \leftarrow {\theta_G}^{(k-1)} - \eta\nabla_{\theta_G}\ell_{\mathcal{X}}(G, H, D_Y, b_X)$;
  ${\theta_H}^{(k)} \leftarrow {\theta_H}^{(k-1)} - \eta\nabla_{\theta_H}\ell_{\mathcal{X}}(G, H, b_X)$;
end
return $G(\cdot;\theta_G)$, $H(\cdot;\theta_H)$, $D_Y(\cdot;\theta_{D_Y})$
`GetLocalGradientFromClient` (θ_{D_Y})**:**
  $b_Y \leftarrow$ (sample batch from client $\mathcal{Y}$) ;
  $g_{\theta_{D_{Y,real}}} \leftarrow \nabla_{\theta_{D_Y}} \ell_{\mathcal{Y}}(D_Y, b_Y)$;
  return $g_{\theta_{D_{Y_{real}}}}$

Algorithm 1 describes the training process of FedCUT. In the first step, the server initializes the parameters of the generator $G(\cdot;\theta_G)$, MLP network $H(\cdot;\theta_H)$, and the discriminator $D_Y(\cdot;\theta_{D_Y})$. For each round k in training, the server sends current parameters of the discriminator $\theta_{D_Y}^{(k-1)}$ to the client. The client then calculates local gradient using its personal data and transmits gradients $g_{\theta_{D_{Y_{real}}}}$ to the server. Next, the server calculates the gradients $g_{\theta_{D_{Y,fake}}}$ for the discriminator using

fake samples. Then, we can update the parameters of the discriminator ${\theta_{D_Y}}^{(k)}$ using $g_{\theta_{D_{Y,real}}}$ and $g_{\theta_{D_{Y,fake}}}$, which is followed by updating parameters of the generator and MLP network: ${\theta_G}^{(k)}$ and ${\theta_H}^{(k)}$. The training round is repeated until the convergence of networks.

A.2 ADDITIONAL RESULTS

We show additional results for natural image translation tasks and low-dose CT denoising task. Fig. 6 shows results for horse-to-zerbra. Fig. 7 shows translation results on cat-to-dog task. Fig. 8 shows results for cityspaces dataset. Fig. 9 shows denoising results for low-dose CT images.



(a) Source (b) CUT (c) FastCUT (d) FedCycleGAN (e) **FedCUT**

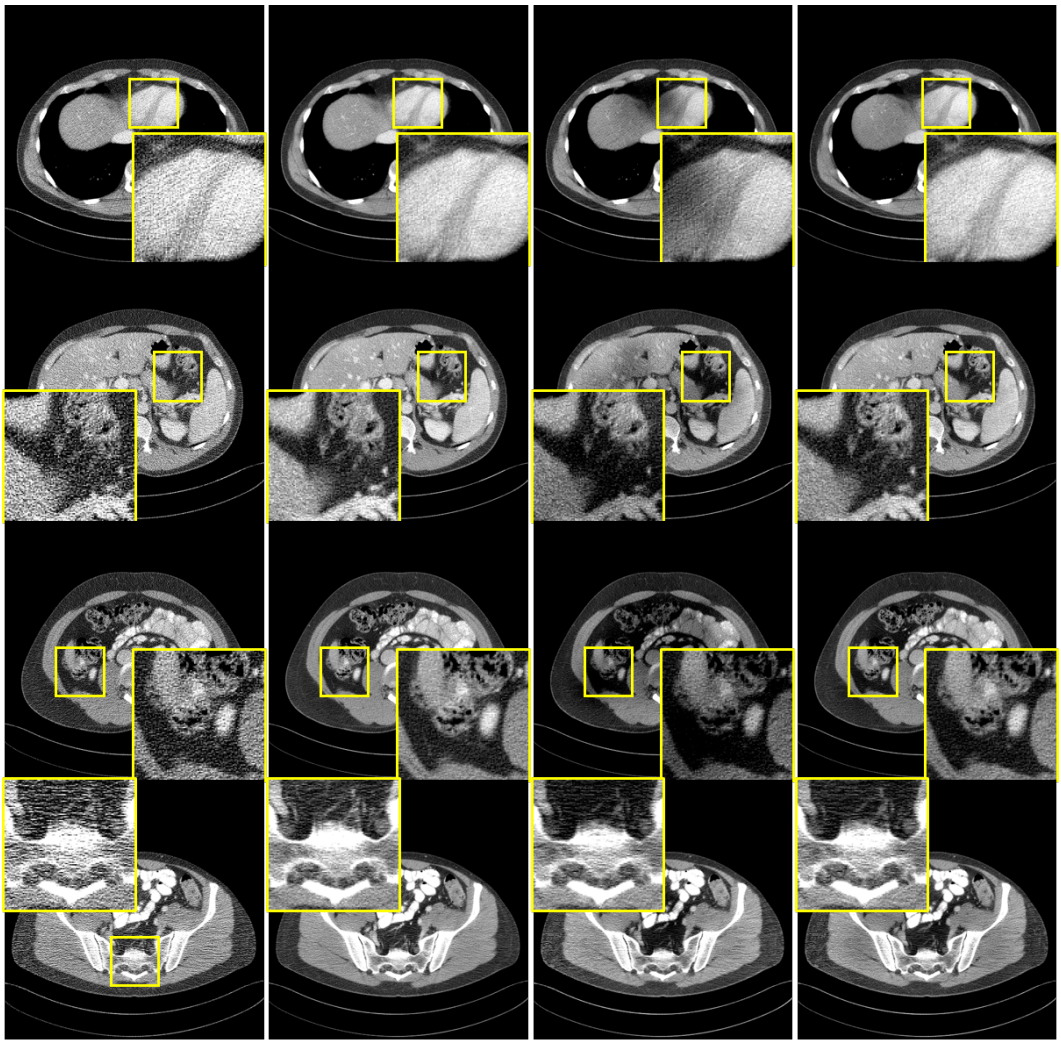
Figure 6: Additional results for horse-to-zebra.

(a) Source (b) CUT (c) FastCUT (d) FedCycleGAN (e) **FedCUT**

Figure 7: Additional results for cat-to-dog.

(a) Source (b) CUT (c) FastCUT (d) FedCycleGAN (e) **FedCUT**

Figure 8: Additional results for Cityscapes.



(a) Low-dose CT (b) Routine-dose CT (c) FedCycleGAN (d) FedCUT

Figure 9: Additional results for low-dose CT denoising. Intensity range is (-160, 240) [HU] (Hounsfield Unit).