ST-FL: Style Transfer Preprocessing in Federated Learning for COVID-19 Segmentation

Antonios Georgiadis[1,*], Varun Babbar[1,2,*], Fran Silavong[1], and Sean Moran[1]

[1]CTO Applied Research: JP Morgan Chase
[2]University of Cambridge
[*]These authors contributed equally to this work.

ABSTRACT

Chest Computational Tomography (CT) scans present low cost, speed and objectivity for COVID-19 diagnosis and deep learning methods have shown great promise in assisting the analysis and interpretation of these images. Most hospitals or countries can train their own models using in-house data, however empirical evidence shows that those models perform poorly when tested on new unseen cases, surfacing the need for coordinated global collaboration. Due to privacy regulations, medical data sharing between hospitals and nations is extremely difficult. We propose a GAN-augmented federated learning model, dubbed ST-FL (Style Transfer Federated Learning), for COVID-19 image segmentation. Federated learning (FL) permits a centralised model to be learned in a secure manner from heterogeneous datasets located in disparate private data silos. We demonstrate that the widely varying data quality on FL client nodes leads to a sub-optimal centralised FL model for COVID-19 chest CT image segmentation. ST-FL is a novel FL framework that is robust in the face of highly variable data quality at client nodes. The robustness is achieved by a denoising CycleGAN model at each client of the federation that maps arbitrary quality images into the same target quality, counteracting the severe data variability evident in real-world FL use-cases. Each client is provided with the target style, which is the same for all clients, and trains their own denoiser. Our qualitative and quantitative results suggest that this FL model performs comparably to, and in some cases better than, a model that has centralised access to all the training data.

Keywords: Style transfer, federated learning, COVID-19 segmentation

1. SUMMARY

We developed a noise-agnostic flavour of Federated Learning by utilizing style transfer prepossessing prior to the federation - the ultimate goal is to achieve better results in COVID-19 segmentation, when the models are trained with FL. The target style is shared in the form of a dataset (*e.g.* 20-50 images), and each client trains its own CycleGAN transformation which maps its local data to the common style. The benefit of this approach is that it reduces any image noise in the CT scans prior to sending information via the federation. We used two publicly available COVID-19 segmentation datasets, in addition to artificial noise patterns, and demonstrated statistically significant lesion segmentation improvement ranging between 5%-90%, depending on the noise pattern.

2. PURPOSE

To design a novel method for achieving noise-agnostic Federated Learning for COVID-19 CT scan segmentation by applying style transfer techniques.

3. INTRODUCTION

The global race against time to tackle and subdue the COVID-19 pandemic has generated an unprecedented level of scientific progress and collaboration worldwide. Nowhere is this progression more apparent in the variety and depth of the research that has and is taking place in image processing and analysis for COVID-19 diagnosis.[1–6] However, due to stringent privacy laws, the sharing of confidential data between institutions and countries is fraught with difficulties, and is generally considered impossible. *Federated Learning* provides a solution to this data sharing dilemma, allowing globally distributed data to remain private while still permitting a centralised neural network model to be learnt using information from all of these images existing across institution and

country boundaries. Federated learning solves the problem of how to learn a single model based on data that is locked away in data silos without revealing per-client private data to other clients or the central server. The client and the aggregator share the same neural network architecture. Clients train on their local data and send the gradient updates to the aggregator, these gradient updates are combined by the aggregator potentially in a cryptographically secure manner,[7] the central model weights are updated with the aggregared gradients, and the resulting weights are distributed to the clients at the same time. In FL, only the model weights are shared between clients and a central server (the aggregator) and not the actual training data, which is considered private and highly confidential.

Prior research has explored the benefits of federated learning for leveraging disparate datasets for the purpose of COVID-19 chest CT scan segmentation.[2] However, there is no previous research that accounts for the differing factors of variation of CT images that are distributed across client nodes. In practice CT images arising from different generations of CT machine can differ vastly across many factors of variation, for example brightness, detail and noise level, in addition to factors such as using a contrast-enhancing agent prior to the scan (contrast-enhanced vs non-contrast images). In our work, we demonstrate that these different factors of variation in the training CT images present significant robustness challenges for federated learning of deep neural networks, leading to sub-optimal models if not properly addressed. Ideally we require a neural network that standardises the training images across all of the client nodes, according to a benchmark image quality. Unfortunately, such a network requires paired training data to learn, and it is not possible to collect samples of low and high quality CT images from the same machine. To address this issue, in this paper, we instead assume that a small representative dataset can be shared with the clients, with the style most commonly encountered, and thus have the clients learn an *unpaired* domain mapping between the local and target domains using a CycleGAN. In our experimental evaluation we demonstrate the benefit of this approach, showing that *ST-FL*, our FL framework incorporating data quality equalisation on the client nodes, leads to significantly more accurate federated models for CT image segmentation, closely approaching the oracle upper-bound of a model learnt with centralised data.

Closely related work to ST-FL is the research of Yang *et al.*,[2] who propose a semi-supervised federated learning framework for chest CT scan segmentation that leverages both labelled and unlabelled data at client nodes and is evaluated over multi-national data from China, Italy and Japan. This work shows the benefit of exploiting unannotated CT scan images in an FL setup for the task of image segmentation. In contrast to our work, they do not address the mixed data issue and the fact that in realistic scenarios the CT images on each client node can vary massively in quality. In other recent related work Jeong *et al.*[8] propose a semi-supervised FL framework that tackles the issue where private client data contains only partial or no labels. Data normalisation is tackled in[9] but they employ fixed transformations *e.g.* Lanczos interpolation, to standardise the hetrogenous client data, whereas we exploit the non-linear mappings possible through deep neural networks.

To summarise, in this paper, our contribution to the state-of-the-art is three-fold:

- **Mixed CT image data quality & the effect on FL**: Through experimentation with synthetic and semi-synthetic datasets of varying structural and stylistic features, we highlight and demonstrate the negative effect of differing quality images on client nodes on the accuracy of a federated U-Net[10] for CT image segmentation.
- **Normalising image quality across client nodes for FL:** We propose two approaches for *normalising* the image quality on client nodes with a CycleGAN.[11] i) *Universal CycleGAN*: only one denoiser is trained at the aggregator level and is then shared with the clients. ii) *Client-specific CycleGAN*: multiple client-specific denoisers are trained at a client level. Both approaches map client domains to a shared common domain, but have different assumptions, pros and cons. This provides the FL framework with similar quality CT images across all client nodes, counteracting the significant domain shift found in practice.
- **Noise agnostic FL framework for different types of noise**: We present *ST-FL*, a federated learning framework that incorporates the denoising CycleGAN at each client node, standardising image quality per client and increasing the robustness of federated learning to mixed data quality observed in practice. Experimental evaluation shows that ST-FL leads to higher quality segmentation models for chest CT scan images.

4. METHODOLOGY

We used a number of publicly available COVID-19 segmentation datasets, which include segmentation masks generated by radiologists. We extracted a small amount of data to be our *target style* and used the rest for training and testing. As the datasets include different cases from around the world, we distributed them in such a way that no client would ever have data from more than one source. In addition, some datasets include multiple slices for the same patient (as CT scans are 3D), so we ensured to split those datasets by patient to avoid cross-client leakage. The goal is to recreate as realistic a scenario as possible, where one hospital would not be able to share data with another.

In addition to the already-existing noise patterns of the dataset (*e.g.* discolouration, blurring, contrast), we further enhanced them with artificial noise (*e.g.* contrast enhancement, contrast inversion, Gaussian noise, mixed noise *etc.*). We experimented with two approaches: i) Universal Cycle-GAN and ii) Client-specific Cycle-GAN, and compared the segmentation results with respect to the FedAvg scheme and a Centralised model. The CycleGANs consist of a U-Net generator and a PatchGAN[12] discriminator. The COVID-19 segmentation model is also a U-Net, which itself is quite effective in ignoring noise[13],[14] however, we still noticed an appreciable improvement in segmentation performance with our proposed framework as compared to FedAvg.

4.1 FedAvg Scheme

The classic Federated Learning setup consists of $k \in \mathcal{Z}^+$ clients, each of which have a local model $\mathcal{G}\omega_k$, parameterised by weights ω_k, that is trained on their local dataset. After every τ epochs, these client weights are transmitted to a central server, where a weighted average, i.e. $\omega_{avg} = \sum_{i=1}^{k} \rho_i \omega_k$, is performed. In practice, the number of clients is generally very large, so only a random subset of clients are sampled and their weights averaged. In our setup, we set $\rho = \frac{1}{k}$ for all experiments. Our experiments are based on the task of binary classification, wherein we are given datasets of CT scans of COVID-19 patients. Our objective is to segment these scans for the presence of lesions. Given the k^{th} client's model prediction for the i^{th} example $\mathcal{G}\omega_k(x_k^i) \in [0, 1]$ and the corresponding ground truth label, $y_k^i \in \{0, 1\}$ our client objective function is a simple cross entropy loss:

$$\mathcal{L}_{CE}(y_k^i, \mathcal{G}\omega_k(x_k^i)) = y_k^i \log(\mathcal{G}\omega_k(x_k^i)) + (1 - y_k^i) \log(1 - \mathcal{G}\omega_k(x_k^i)) \tag{1}$$

For segmentation, this loss is applied pixelwise and averaged over the image.

4.2 CycleGAN Based PreProcessing

In recent years, the CycleGAN paradigm developed by Zhu *et al.*[11] has shown great promise in mapping data sets between different styles. This scheme consists of 2 complementary generators $\mathcal{G}_{\mathcal{AB}}(x_A) : A \to B$ and $\mathcal{G}_{\mathcal{BA}}(x_B) : B \to A$ for style domains A and B and examples x_A and x_B. Each generator is coupled with associated adversarial discriminators $D_A(x_A||\mathcal{G}_{\mathcal{BA}}(x_B))$ and $D_B(x_B||\mathcal{G}_{\mathcal{AB}}(x_A))$. We aim to solve the following optimisation problem:

$$\mathcal{G}_{BA}^*, \mathcal{G}_{AB}^* = \min_{\mathcal{G}_{BA}, \mathcal{G}_{AB}} \max_{D_A, D_B} \mathcal{L}(\mathcal{G}_{BA}, G_{AB}, D_B, D_A, x_A, x_B) \tag{2}$$

where

$$\mathcal{L}(\mathcal{G}_{BA}, \mathcal{G}_{AB}, D_B, D_A, x_A, x_B) = \mathcal{L}_{GAN}(\mathcal{G}_{BA}, D_A, x_B, x_A) + \mathcal{L}_{GAN}(\mathcal{G}_{AB}, D_B, x_B, x_A) + \lambda \mathcal{L}_{Cyc}(\mathcal{G}_{BA}, G_{AB}, x_B, x_A) \tag{3}$$

where the first 2 losses in equation 3 are adversarial losses for the 2 complementary GANs. They take the form:

$$\mathcal{L}_{GAN}(\mathcal{G}_{BA}, D_A, x_B, x_A) = \mathbb{E}_{x_A \sim p_{data}(x_A)}[\log(D_A(x_A)] + \mathbb{E}_{x_B \sim p_{data}(x_B)}[\log(1 - D_A(\mathcal{G}_{BA}(x_B)))] \tag{4}$$

The third loss, called the cycle consistency loss, is a regulariser used to enforce translation back to the original image when the style transfer generators are applied in succession. Concretely, we want $G_{BA}(G_{AB}(x_A)) \approx x_A$ and $G_{AB}(G_{BA}(x_B)) \approx x_B$. Thus:

$$\mathcal{L}_{cyc} = \mathbb{E}_{x_A \sim p_{data}(x_A)}[||\mathcal{G}_{BA}(\mathcal{G}_{AB}(x_A)) - x_A||_1] + \mathbb{E}_{x_B \sim p_{data}(x_B)}[||\mathcal{G}_{AB}(\mathcal{G}_{BA}(x_B)) - x_B||_1] \tag{5}$$

In this paper, we consider the scenario where clients have access to a common dataset that represents a target style and propose 2 methods: Universal CycleGAN style transfer and Client Specific CycleGAN style transfer. While these approaches would be task-dependant and entail overhead costs of extra training, these would be one time costs, and fast style transfer is possible once models are trained.

4.2.1 Universal CycleGAN

In this approach, we train a CycleGAN on an aggregated dataset $D_{agg} = \cup_{k=1}^{N} \mathcal{D}_{sub}^{k}$ where $D_{sub}^{k} \in \mathcal{X}_k \times Y_k$ is a random subset of the k^{th} client's dataset D^k. To ensure equal representation of all client styles, we fix $|D_{sub}^{k}| = 25 \; \forall \; k$ in our experiments. Figure 1 shows a schematic of the proposed setup.

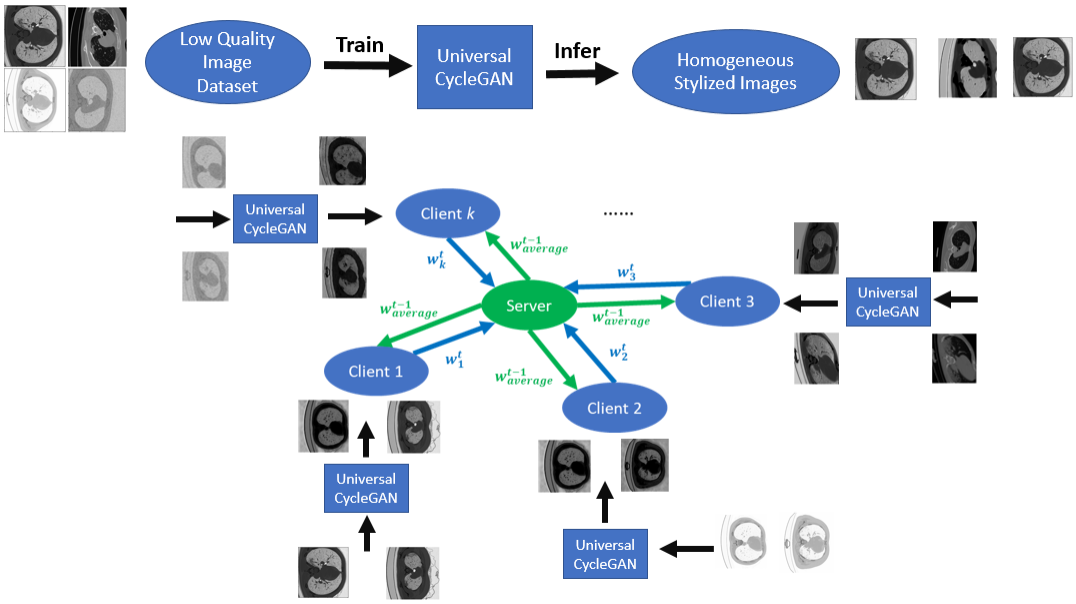

Figure 1: The Universal CycleGAN Scheme for Federated Learning

4.2.2 Client-Specific CycleGAN

In this approach, we train an individual CycleGAN for each client. Specifically, for a client style domain C_k and target domain T (where T is the same for all clients), we learn generators - $\mathcal{G}_{C_k T}(x_{C_k}) : C_k \rightarrow T$ and $\mathcal{G}_{TC_k}(x_T) : T \rightarrow C_k$ - and discriminators $D_{C_k}(x_{C_k} || \mathcal{G}_{TC_k}(x_T))$ and $D_T(x_T || \mathcal{G}_{\mathcal{C}_{||}\mathcal{T}}(x_{C_k}))$. As above, this system is trained using the objective function in Equation 3. Figure 2 shows a schematic of the proposed setup.

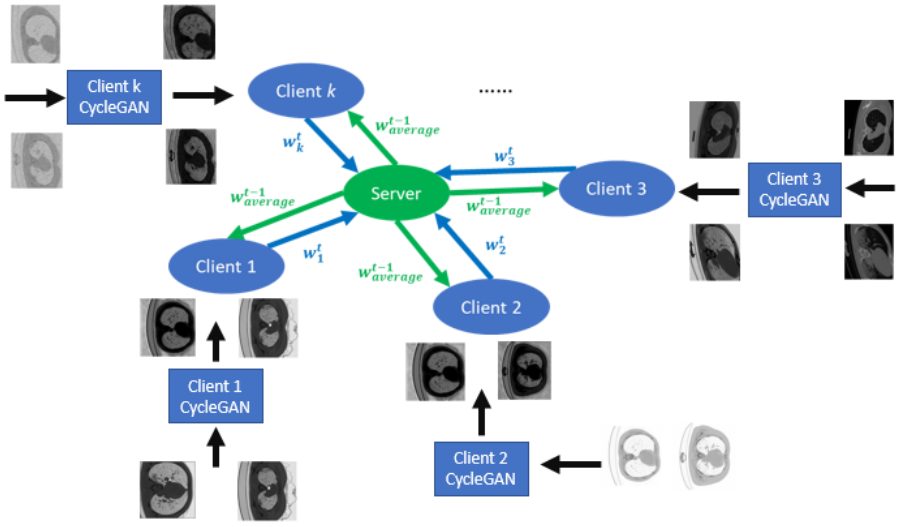

Figure 2: The Client Specific CycleGAN Scheme for Federated Learning

4.3 Training

Prior to the actual federated training, we train both Universal and Client Specific CycleGANs for 100 epochs. For performing federated training, our initial approach had been to supply only the style transferred image. However, this showed limited gains in segmentation performance: this is likely because the process of style transfer introduces noise in the image, leading to some loss of salient information. Henceforth, for our experiments, we concatenate original and style transferred images for each client Segmentation UNet, ensuring that client models can learn salient information from each channel. This ensures that for all schemes, the segmentation performance should be at least as good as FedAvg, on average (because in the worst case scenario, the model weights can learn purely from the original image channel). For client datasets that serve as style targets, we concatenate 2 copies of the same image to serve as inputs. We then train these models in a federated setting for 35 epochs. At the end of each training epoch, we aggregate their weights in a server model and broadcast them back to each client. The entire training process is shown in Algorithms 1, 2, and 3 in the Appendix. For comparative purposes, we also train a centralised model on a pooled dataset $\cup_{k=1}^{N} \mathcal{D}_{train}^{k}$ that has the same 2 channel input as described above.

5. EXPERIMENTS & RESULTS

5.1 Datasets

In order to test the efficacy of our scheme, we perform experiments with 2 different types of client datasets:

- **Synthetic Dataset**: We use the Coronacases[15] dataset of COVID-19 patient chest scans, both in its vanilla form and in an augmented form wherein each client dataset represents different noise patterns added to the dataset. This scenario models a situation where different client institutions may have chest scans with similar structural characteristics but differing style characteristics (eg. differing imaging modalities and noise levels). Figure 3 shows some samples from all client synthetic datasets.
- **Semi-Synthetic Dataset**: We use the Coronacases and MedSeg[16] dataset as client datasets and augment them with similar noise patterns as above to create additional client datasets. The MedSeg dataset originally had labels corresponding to different abnormalities observed in CT scans (ground glass, pleural effusions, and consolidations). Here, we discard all labels except ground glass, which we consider as the binary segmentation target. This was empirically seen to resemble lesions seen in the Coronacases dataset (see Appendix). Compared to the Coronacases dataset, the MedSeg dataset was seen to have noisy labels and some structural differences that can potentially hinder effective training. These data-centric properties model the scenario where client institutions may have scans with differing structural and stylistic characteristics and some improperly labelled examples. We show some generated and real samples for the semi-synthetic dataset scenario in Figure 4.

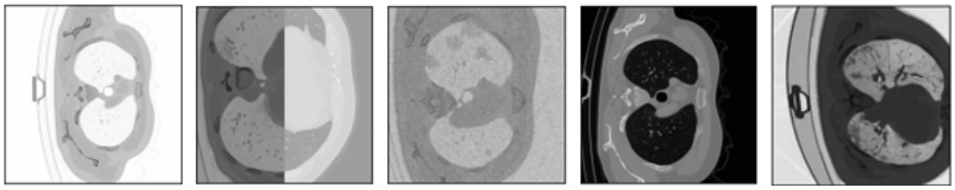

Figure 3: Sample Images from Synthetic Datasets: (Left to Right: Vanilla Coronacases, Mixed Noise Coronacases, Noisy Coronacases, Inversion Coronacases, Contrast Enhanced Coronacases - Style Target)

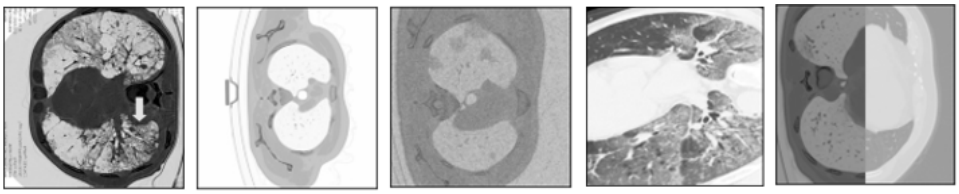

Figure 4: Sample Images from Semi-Synthetic Datasets: (Left to Right: Vanilla Medseg - Style Target, Vanilla Coronacases, Noisy Coronacases, Inversion MedSeg, Mixed Noise Coronacases)

In this paper, we consider the scenario where client datasets are of similar size. Because the Coronacases and MedSeg datasets are of different sizes (30 and 100 images respectively), we fix $|\mathcal{D}^k| = 30$ for all clients. For both approaches, we add random warping to all client images not only for data augmentation, but also to add some variability in different client datasets. This also ensures that the CycleGAN is able to learn unpaired mappings between the original and target styles and becomes agnostic to any structural similarity between images. To warp images, we apply a random 4-point perspective transform[17] to the images, with each of the 4 points sampled from a normal distribution. The variance of this distribution is uniformly sampled in the interval $[0.1, 0.15]$. After warping, we keep aside 20% of each client dataset $\mathcal{D}^k_{val}$ as a validation set and calculate the average Dice and IOU score on the union of all client validation sets $\mathcal{D}_{\text{val}} = \cup_{k=1}^{N} \mathcal{D}^k_{val}$. To understand which noise patterns the CycleGAN based approaches perform well on, we calculate the average performance improvement on the union of training and validation sets for each client.

For our scheme, we assume that each client has access to a task-dependent exemplar dataset that embodies a universal style target. Our experimentation suggests the following guidelines for a suitable style target:

- If there are datasets with structural dissimilarities but similar style, the style target should ideally contain examples of these structurally dissimilar clients. For example, in the semi-synthetic dataset scenaro, we chose the style target as a random subset of the MedSeg and Coronacases Contrast datasets (which are stylistically similar), as all other client datasets are derived from these datasets, albeit with added warping and noise patterns.
- The style target should contain images that are stylistically homogeneous but distinct from the style of the client datasets. Figures 3 and 4 show distinct client images of varying noise patterns and the corresponding style target. Having stylistically distinct client and target images ensures a more well defined mapping between the original and desired style domain.

5.2 Results

Table 1 shows metric scores of the different schemes tested across clients and dataset types. In each case, we report the highest recorded Dice and Intersection Over Union (IOU) scores for thresholded segmentation $0 - 1$ masks (with threshold probability 0.35) averaged over 5 trials and their associated 95 % confidence intervals. We also report the validation Dice and IOU scores for all schemes averaged across number of clients, client datasets, and 5 trials in Figure 5. Note that the initial plateau occurs because we report hard dice scores, making initial model outputs look the same after thresholding. However, the analysis and results were seen to remain the same for soft scores as well. For all dataset types and number of clients, we observe that the client specific CycleGAN preprocessing scheme outperforms its federated learning counterparts in both metrics. For synthetic datasets with high degree of structural similarity, the centralised training scheme can be viewed as an upper bound on the segmentation performance relative to all other federated learning schemes, with the client specific CycleGAN scheme coming closest to this bound. For semi-synthetic datasets, on the other hand, we observe no discernable pattern in segmentation performance of centralised training, with the client specific CycleGAN scheme outperforming it in all cases. Intuitively, this is likely because the centralised model is being trained on datasets of varying structural similarities and noisy labels, making it harder for weights to generalise across datasets. The client specifc CycleGAN becomes more robust, leading to improved performance across all noise patterns tested. We also see that universal CycleGAN preprocessing offers lower and more inconsistent

performance gains on average compared to the client specifc CycleGAN scheme. This is because the style transfer is often of very poor quality, especially in situations involving large numbers of clients, as the model is unable to adapt to differing client distributions (see Figure 7 in the Appendix).

Number of Clients	Metric	Dataset Type	Federated Training			Centralised Training
			Vanilla FedAvg	Universal CycleGAN	Client Specific CycleGAN	
3	Dice	Synthetic	0.414 ±0.012	0.505 ±0.027	**0.533 ±0.041**	0.560 ±0.023
		Semi-Synthetic	0.497 ±0.009	0.520 ±0.012	**0.539 ±0.012**	0.494 ±0.009
	IOU	Synthetic	0.274 ±0.011	0.329 ±0.017	**0.336 ±0.019**	0.347 ±0.016
		Semi-Synthetic	0.337 ±0.014	0.355 ±0.016	**0.366 ±0.006**	0.338 ±0.010
4	Dice	Synthetic	0.430 ±0.008	0.493 ±0.026	**0.514 ±0.037**	0.528 ±0.017
		Semi-Synthetic	0.462 ±0.014	0.483 ±0.021	**0.489 ±0.014**	0.450 ±0.013
	IOU	Synthetic	0.287 ±0.007	0.296 ±0.013	**0.332 ±0.017**	0.329 ±0.007
		Semi-Synthetic	0.306 ±0.007	0.319 ±0.006	**0.321 ±0.004**	0.305 ±0.008
5	Dice	Synthetic	0.378 ±0.012	0.460 ±0.025	**0.465 ±0.006**	0.490 ±0.006
		Semi-Synthetic	0.390 ±0.018	0.420 ±0.017	**0.462 ±0.017**	0.392 ±0.018
	IOU	Synthetic	0.255 ±0.008	0.240 ±0.009	**0.282 ±0.013**	0.310 ±0.012
		Semi-Synthetic	0.265 ±0.007	0.281 ±0.012	**0.301 ±0.010**	0.261 ±0.010

Table 1: Best Case Federated and Centralised Performance Metrics for Differing Numbers of Clients. We report the best metric averaged over 5 trials and its associated 95% CI.

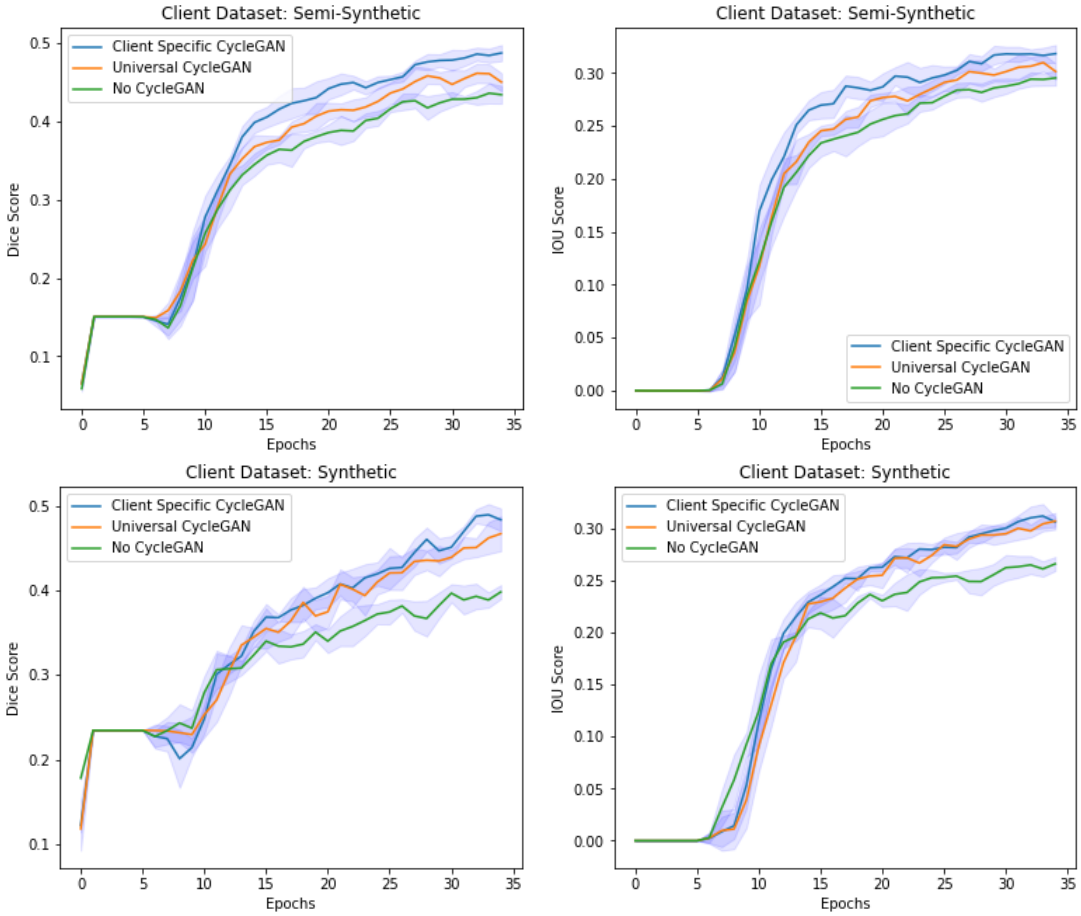

Figure 5: Thresholded Validation Dice / IOU Curve of Federated Schemes Averaged Across All Experiments, Noise Patterns, and 5 Trials. (95% CI overlaid in grey)

Figure 6 shows a bar plot of the performance gains of all the scheme tested relative to the vanilla FedAvg scheme. Here, we averaged the performance improvement of each noise pattern across differing numbers of clients tested. As such, we report this % improvement averaged over 5 runs. We note that there is a discrepancy in the performance of CycleGAN related schemes over the noise patterns tested. Specifically, for both the semi synthetic and synthetic datasets, applying style transfer preprocessing on images corrupted by Gaussian noise does not produce meaningful improvement in dice scores. This is likely due to information loss in images that is not necessarily corrected by style transfer. Conversely, we see significant gains in performance in inversion and mixed noise patterns for both dataset types, though these gains are larger for the synthetic dataset case. Intuitively, this is because averaging weights of models trained on structurally dissimilar client datasets likely limits the benefits of style transfer, which can only correct for noise distribution shifts and not structural shifts.

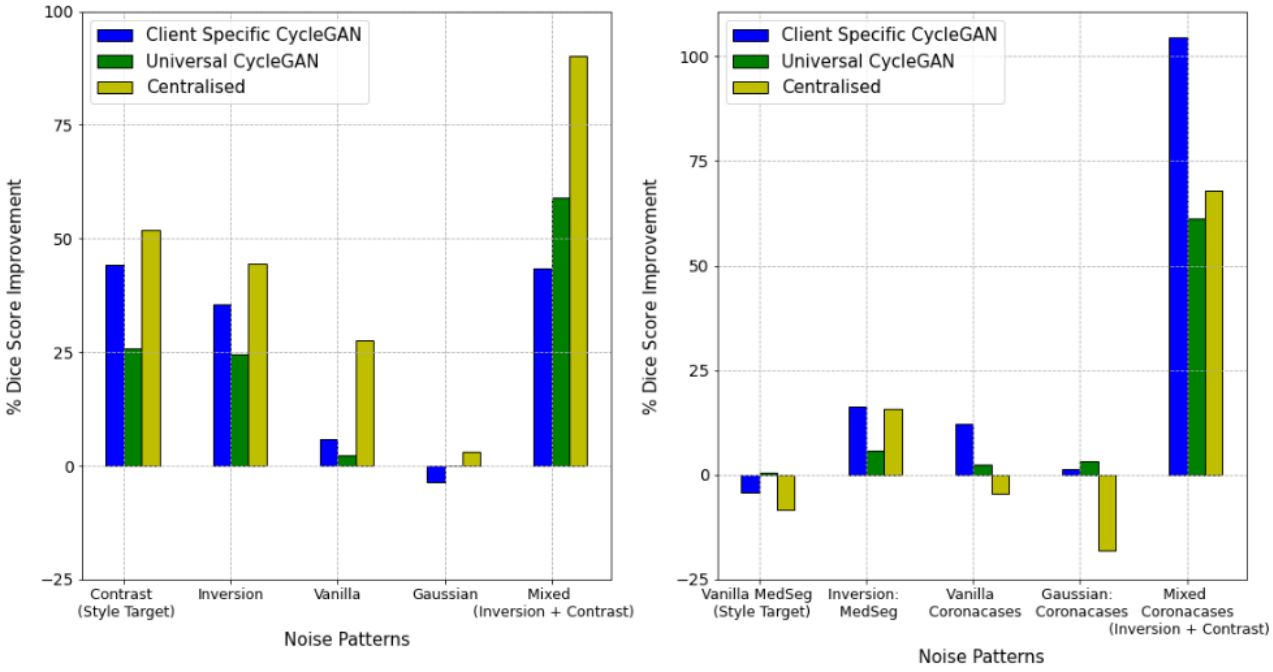

Figure 6: Best Case % Noise-Specific Performance Improvement of the Different Techniques Tested Relative to Vanilla FedAvg (Left: Synthetic Datasets, Right: Semi-Synthetic Datasets)

6. NOVEL WORK & CONCLUSIONS

This paper presents a novel preprocessing method for performing federated learning in a noise agnostic manner, with a focus on segmentation of lesions in COVID-19 patient chest scans. Medical datasets in a federated learning setup tend to have variations in contrast, noise, brightness and detail, motivating the need for a common normalisation scheme which renders federated systems agnostic to noise. We explored the idea of using style transfer based pre-processing on client datasets in 2 scenarios: a) varying noise patterns but common structure, and b) varying noise patterns and varying structure. Our work suggests that style transfer pre-processing leads to higher dice scores in downstream segmentation tasks on average in both cases. We further characterised the performance of our method on some common noise patterns in medical datasets and found disparities, with some noise patterns showing much more improvement in segmentation performance than others.

Future work could focus on exploration of this technique in settings where client datasets are unbalanced and / or are of unequal size. There is also a need to further characterise noise-specific performance and explore other style transfer techniques that could be potentially useful.

REFERENCES

[1] Sheller, M., Edwards, B., Reina, G., Martin, J., Pati, S., Kotrotsou, A., Milchenko, M., Xu, W., Marcus, D., Colen, R., and Bakas, S., "Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data," *Scientific Reports* **10** (Dec. 2020). Publisher Copyright: © 2020, The Author(s). Copyright: Copyright 2020 Elsevier B.V., All rights reserved.

[2] Yang, D., Xu, Z., Li, W., Myronenko, A., Roth, H. R., Harmon, S., Xu, S., Turkbey, B., Turkbey, E., Wang, X., et al., "Federated semi-supervised learning for covid region segmentation in chest ct using multi-national data from china, italy, japan," *Medical image analysis* **70**, 101992 (2021).

[3] Sandfort, V., Yan, K., Pickhardt, P., and Summers, R., "Data augmentation using generative adversarial networks (cyclegan) to improve generalizability in ct segmentation tasks," *Scientific Reports* **9** (2019).

[4] Yin, Z., Xia, K., He, Z., Zhang, J., Wang, S., and Zu, B., "Unpaired image denoising via wasserstein gan in low-dose ct image with multi-perceptual loss and fidelity loss," *Symmetry* **13**(1), 126 (2021).

[5] Waheed, A., Goyal, M., Gupta, D., Khanna, A., Al-Turjman, F., and Pinheiro, P. R., "Covidgan: Data augmentation using auxiliary classifier gan for improved covid-19 detection," *IEEE Access* **8**, 91916–91923 (2020).

[6] Laradji, I., Rodriguez, P., Manas, O., Lensink, K., Law, M., Kurzman, L., Parker, W., Vazquez, D., and Nowrouzezahrai, D., "A weakly supervised consistency-based learning method for covid-19 segmentation in ct images," in [*Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*], 2453–2462 (January 2021).

[7] Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., Ramage, D., Segal, A., and Seth, K., "Practical secure aggregation for privacy-preserving machine learning," in [*Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security*], *CCS '17*, 1175–1191, Association for Computing Machinery, New York, NY, USA (2017).

[8] Jeong, W., Yoon, J., Yang, E., and Hwang, S. J., "Federated semi-supervised learning with inter-client consistency," (2020).

[9] Kumar, R., Khan, A. A., Zhang, S., Wang, W., Abuidris, Y., Amin, W., and Kumar, J., "Blockchain-federated-learning and deep learning models for COVID-19 detection using CT imaging," *CoRR* **abs/2007.06537** (2020).

[10] Ronneberger, O., Fischer, P., and Brox, T., "U-net: Convolutional networks for biomedical image segmentation," in [*Medical Image Computing and Computer-Assisted Intervention 2015*], Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F., eds., 234–241, Springer International Publishing, Cham (2015).

[11] Zhu, J.-Y., Park, T., Isola, P., and Efros, A. A., "Unpaired image-to-image translation using cycle-consistent adversarial networks," in [*Proceedings of the IEEE international conference on computer vision*], 2223–2232 (2017).

[12] Isola, P., Zhu, J.-Y., Zhou, T., and Efros, A. A., "Image-to-image translation with conditional adversarial networks," (2018).

[13] Vojtekova, A., Lieu, M., Valtchanov, I., Altieri, B., Old, L., Chen, Q., and Hroch, F., "Learning to denoise astronomical images with u-nets," *Monthly Notices of the Royal Astronomical Society* **503**, 3204–3215 (Nov 2020).

[14] Weigert, M., Schmidt, U., Boothe, T., Müller, A., Dibrov, A., Jain, A., Wilhelm, B., Schmidt, D., Broaddus, C., Culley, S., Rocha-Martins, M., Segovia-Miranda, F., Norden, C., Henriques, R., Zerial, M., Solimena, M., Rink, J., Tomancak, P., Royer, L., Jug, F., and Myers, E. W., "Content-aware image restoration: pushing the limits of fluorescence microscopy," *Nature Methods* **15**, 1090–1097 (Nov. 2018).

[15] "Coronacases.org." coronacases.org.

[16] "Covid-19 CT scan image data and segmentation dataset. Free to download." http://medicalsegmentation.com/covid19/.

[17] Horaud, R., Conio, B., Leboulleux, O., and Lacolle, B., "An analytic solution for the perspective 4-point problem," *Computer Vision, Graphics, and Image Processing* **47**, 33–44 (July 1989).

APPENDIX A. ALGORITHM FOR STYLE-TRANSFER FEDERATED LEARNING

Algorithm 1 Style Transfer Pre-Processing: Client Specific CycleGAN

Style Transfer Pre-Processing: Client Specific CycleGAN

Require: Number of Clients $k \in \mathcal{Z}^+$, Style Transfer Training Epochs S, Client Generators and Discriminators $\mathcal{G}_{C_kT}, \mathcal{G}_{TC_k}, \mathcal{D}_{C_k}, \mathcal{D}_T$ (with weights $\omega_{(.)}$), Client Datasets D_k, Style Target Dataset T

for each client $i = 1, 2, ..., k$ **do**
  for each epoch $s = 1, 2, ..., S$ **do**
    for $x_i, y_i \in (D_i, T)$ **do**
      $\omega_{\mathcal{G}_{TC_i}} \rightarrow \omega_{\mathcal{G}_{TC_i}} - \eta \nabla_{\omega_{\mathcal{G}_{TC_i}}} (\mathcal{L}_{GAN}(G_{TC_i}, D_{C_i}, x_i, y_i) + \lambda \mathcal{L}_{Cyc}(G_{TC_i}, G_{C_iT}, x_i, y_i))$
      $\omega_{\mathcal{G}_{C_iT}} \rightarrow \omega_{\mathcal{G}_{C_iT}} - \eta \nabla_{\omega_{\mathcal{G}_{C_iT}}} (\mathcal{L}_{GAN}(G_{C_iT}, D_T, x_i, y_i) + \lambda \mathcal{L}_{Cyc}(G_{TC_i}, G_{C_iT}, x_i, y_i))$
      ▷ Minimize these Losses
      if s is a multiple of 10 **then**
        $\omega_{\mathcal{D}_T} \rightarrow \omega_{\mathcal{D}_T} + \eta \nabla_{\omega_{\mathcal{D}_T}} (\mathcal{L}_{GAN}(G_{C_iT}, D_T, x_i, y_i))$
        $\omega_{\mathcal{D}_{C_i}} \rightarrow \omega_{\mathcal{D}_{C_i}} + \eta \nabla_{\omega_{\mathcal{D}_{C_i}}} (\mathcal{L}_{GAN}(G_{TC_i}, D_{C_i}, x_i, y_i))$ ▷ Maximise these Losses
      end if
    end for
  end for
  Concatenate original and style transferred images channel-wise: $D_i^T = \{(x_i, \mathcal{G}_{C_iT}(x_i)) : x_i \in D_i\}$
end for

Algorithm 2 Style Transfer Pre-Processing: Universal CycleGAN

Style Transfer Pre-Processing: Universal CycleGAN

Require: Number of Clients $k \in \mathcal{Z}^+$, Style Transfer Training Epochs S, CycleGAN Generators and Discriminators $\mathcal{G}_{LQT}, \mathcal{G}_{TLQ}, \mathcal{D}_{LQ}, \mathcal{D}_T$ (with weights $\omega_{(.)}$), Dataset of Low Quality Images $D_{LQ} \subseteq \bigcup_{i=1}^{k} D_i$ for client datasets D_i, Style Target Dataset T

for each epoch $s = 1, 2, ..., S$ **do**
  for $x_i, y_i \in (D_{LQ}, T)$ **do**
    $\omega_{\mathcal{G}_{TLQ}} \rightarrow \omega_{\mathcal{G}_{TLQ}} - \eta \nabla_{\omega_{\mathcal{G}_{TLQ}}} (\mathcal{L}_{GAN}(G_{TLQ}, D_{LQ}, x_i, y_i) + \lambda \mathcal{L}_{Cyc}(G_{TLQ}, G_{LQT}, x_i, y_i))$
    $\omega_{\mathcal{G}_{LQT}} \rightarrow \omega_{\mathcal{G}_{LQT}} - \eta \nabla_{\omega_{\mathcal{G}_{LQT}}} (\mathcal{L}_{GAN}(G_{LQT}, D_T, x_i, y_i) + \lambda \mathcal{L}_{Cyc}(G_{TLQ}, G_{LQT}, x_i, y_i))$
    ▷ Minimize these Losses
    if s is a multiple of 10 **then**
      $\omega_{\mathcal{D}_T} \rightarrow \omega_{\mathcal{D}_T} + \eta \nabla_{\omega_{\mathcal{D}_T}} (\mathcal{L}_{GAN}(G_{LQT}, D_T, x_i, y_i))$
      $\omega_{\mathcal{D}_{LQ}} \rightarrow \omega_{\mathcal{D}_{LQ}} + \eta \nabla_{\omega_{\mathcal{D}_{LQ}}} (\mathcal{L}_{GAN}(G_{TLQ}, D_{LQ}, x_i, y_i))$ ▷ Maximise these Losses
    end if
  end for
end for
for each client $i = 1, 2, ..., k$ **do**
  Concatenate original and style transferred images channel-wise: $D_i^T = \{(x_i, \mathcal{G}_{LQT}(x_i)) : x_i \in D_i\}$

Algorithm 3 Style Transfer FedAvg

Federated Learning

Require: Number of Clients $k \in \mathcal{Z}^+$, Federated Training Rounds N, Client Segmentation UNet G_{ω_k}, Client Style Transferred Datasets $\mathcal{D}_k^T$

for each round $t = 1, 2, ..., N$ **do**
  for each client $i = 1, 2, ..., k$ **do**
    for batch $\mathcal{B} = (x_i^b, y_i^b) \in \mathcal{D}_i^T$ **do**
      $\omega_i^t \rightarrow \omega_i^t - \eta \nabla_{\omega_i^t}(\mathcal{L}_{CE}(y_i^b, G_{\omega_i^t}(x_i^b)))$
    end for
  end for
  $\omega^{t+1} = \frac{1}{k}\sum_{i=1}^{k} \omega_i^t$ ▷ Server Aggregates Weights at the End of Each Round
  Server broadcasts ω^{t+1} back to all clients, i.e. $\omega_i^{t+1} = \omega^{t+1}\ \forall i \in \{1, 2, ..., k\}$
end for=0

APPENDIX B. SEGMENTATION AND STYLE TRANSFER: QUALITATIVE STUDY

Here we perform a qualitative analysis of the segmentation performance of our final techniques relative to FedAvg and Centralised Training.

- We show that applying style transfer pre-processing leads to better segmentation of lesions on average. Our qualitative results are also consistent with noise-specific performance improvements seen in Figure 6.
- We also show that more realistic style transfer takes place when a Client Specific CycleGAN is used as a preprocessor compared to a Universal CycleGAN.

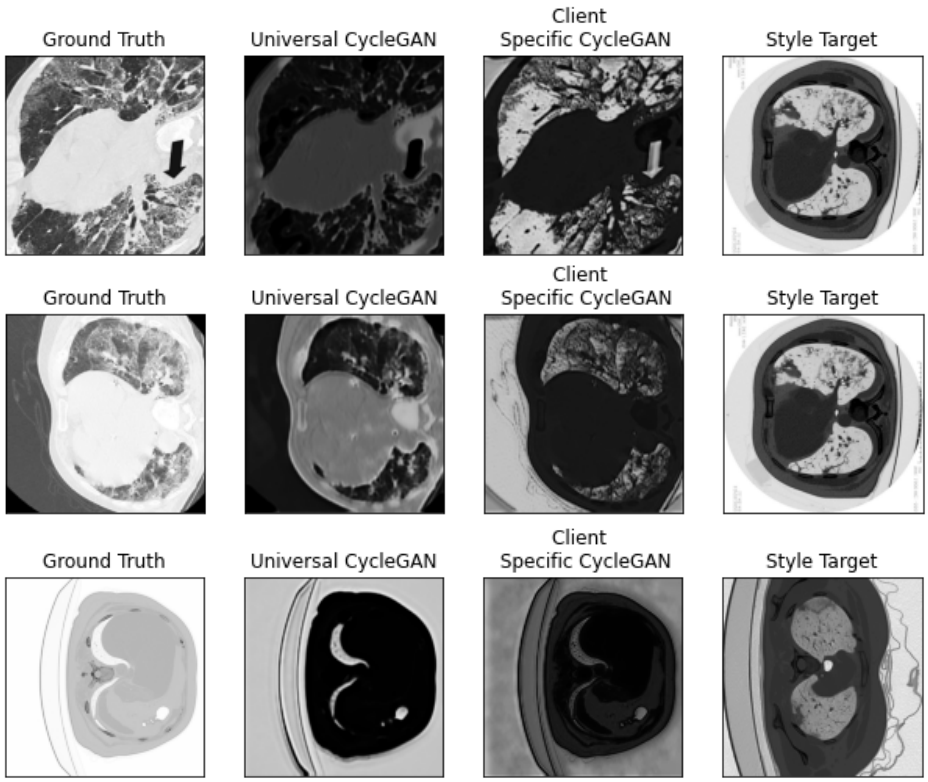

Figure 7: Comparison Between Style Transferred Images for Cycle-GAN Based Pre-Processors

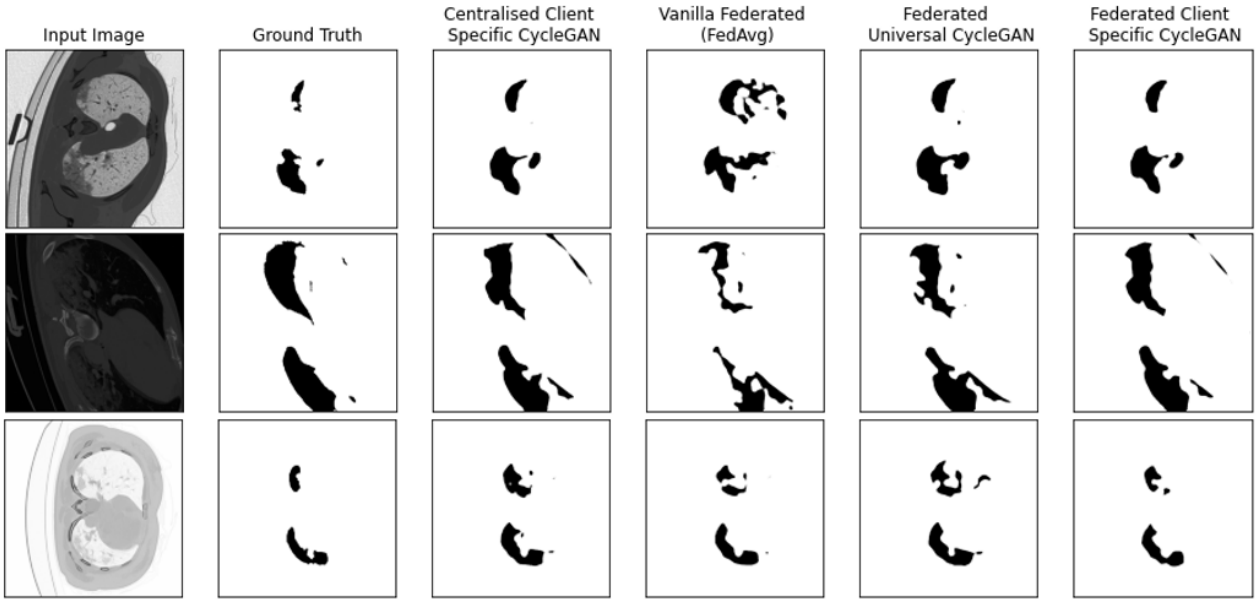

Figure 8: Sample Segmentations of Synthetic Dataset: 3 Clients

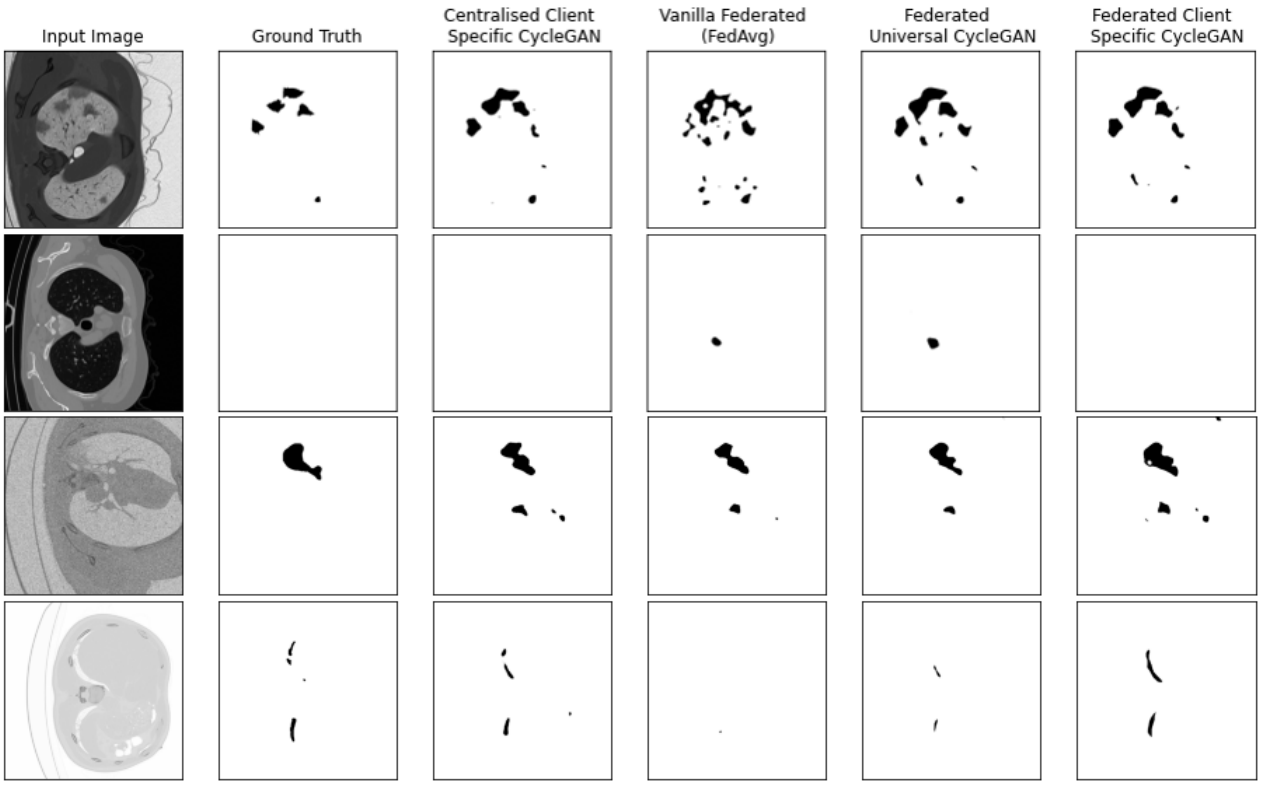

Figure 9: Sample Segmentations of Synthetic Dataset: 4 Clients

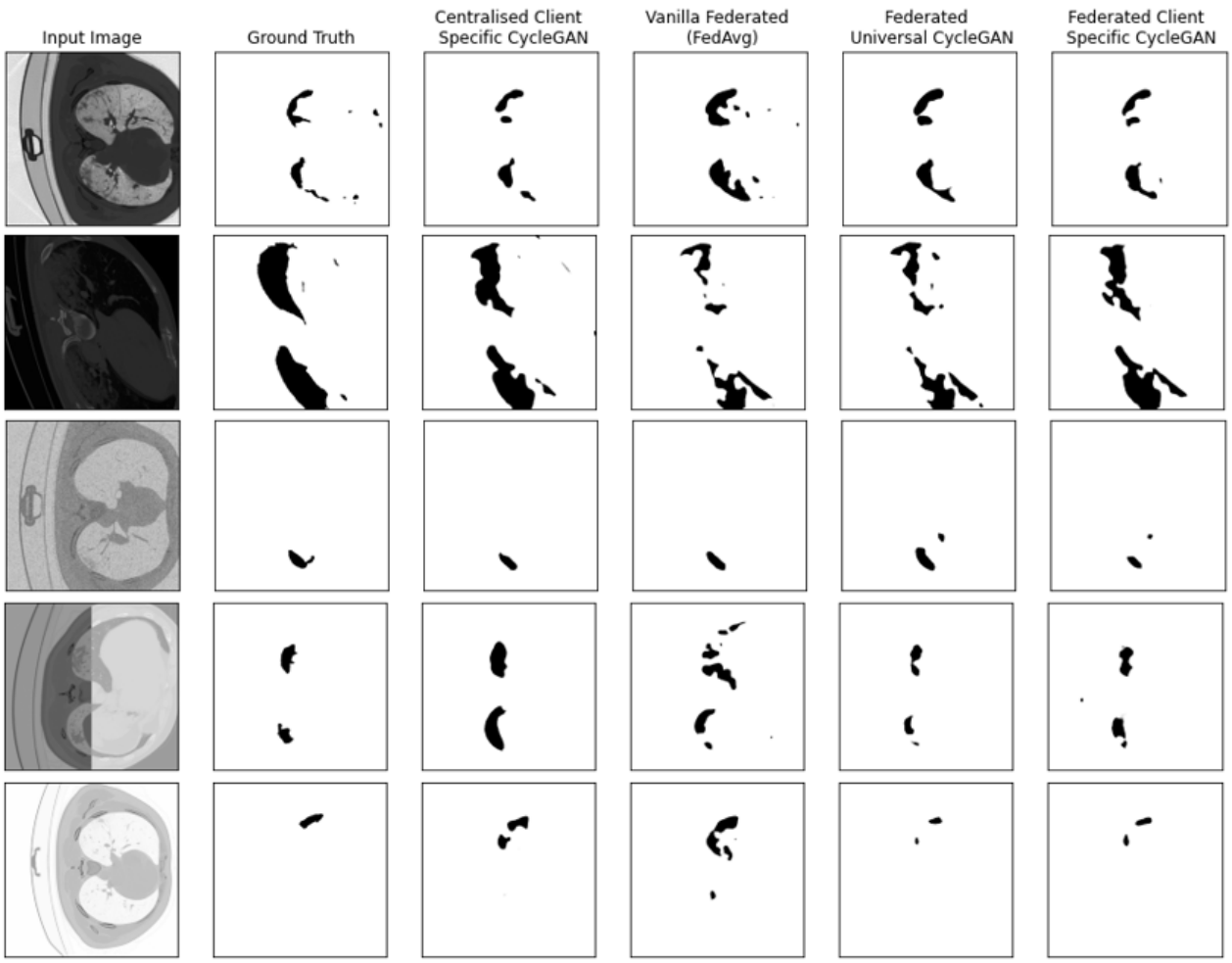

Figure 10: Sample Segmentations of Synthetic Dataset: 5 Clients

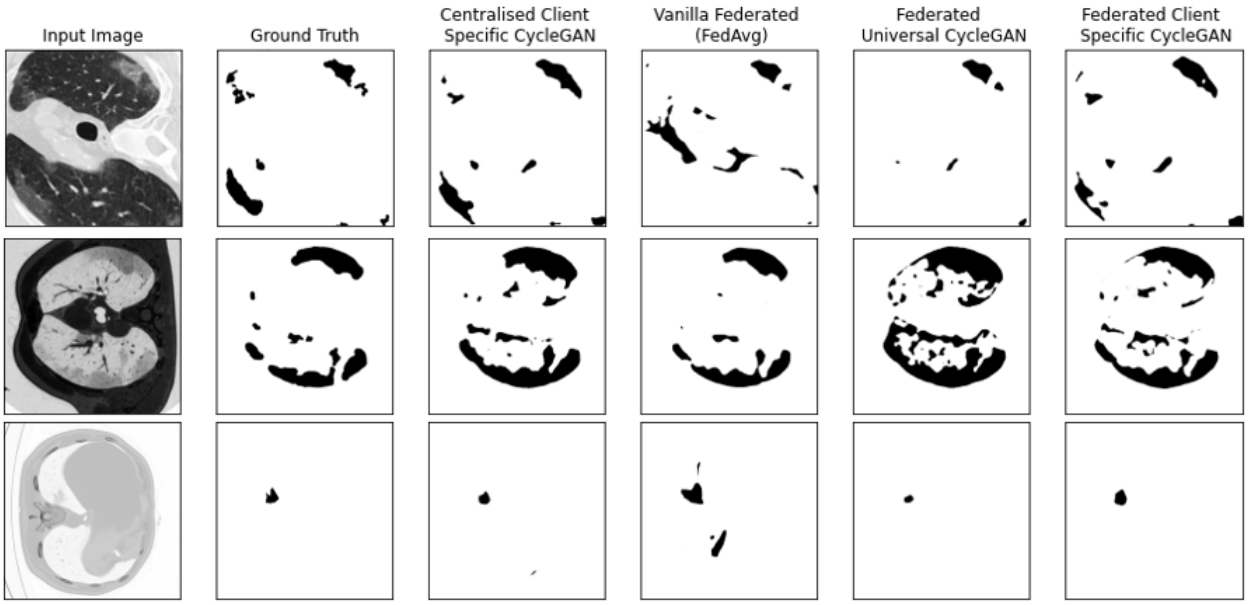

Figure 11: Sample Segmentations of Semi-Synthetic Dataset: 3 Clients

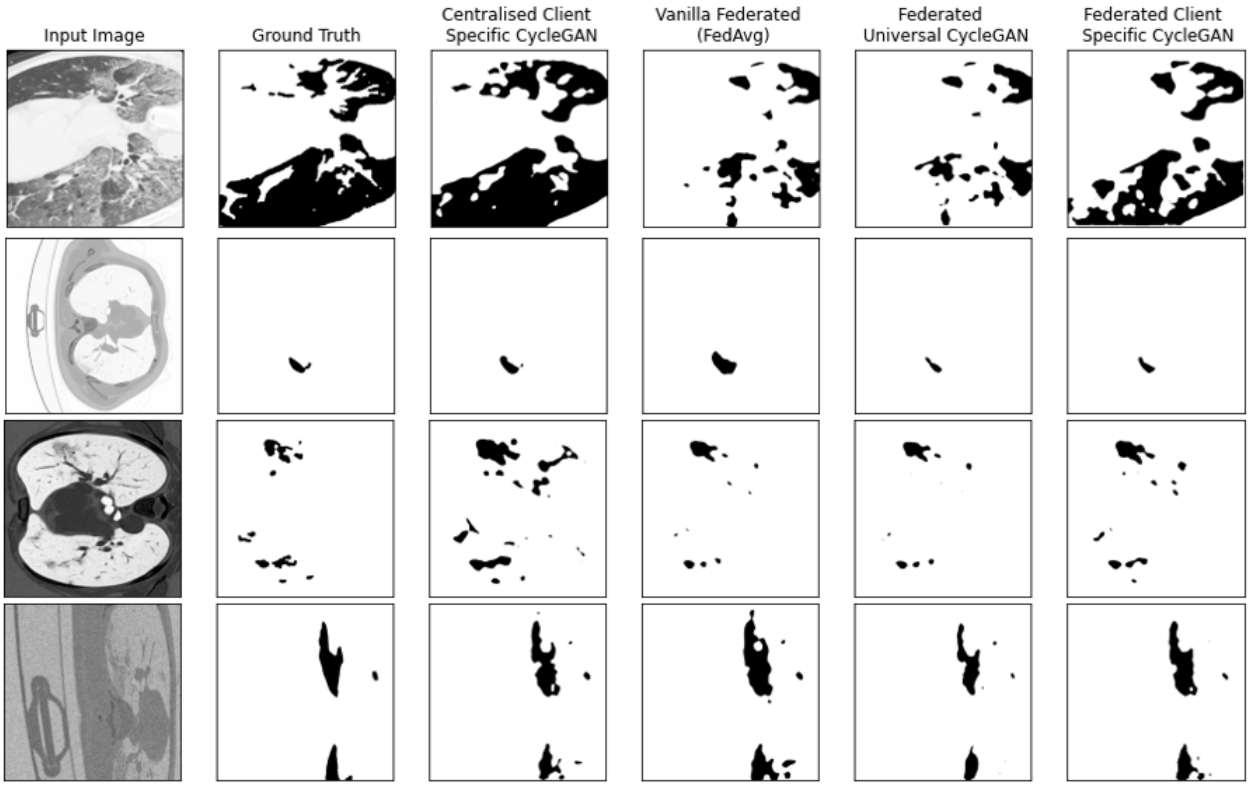

Figure 12: Sample Segmentations of Semi-Synthetic Dataset: 4 Clients

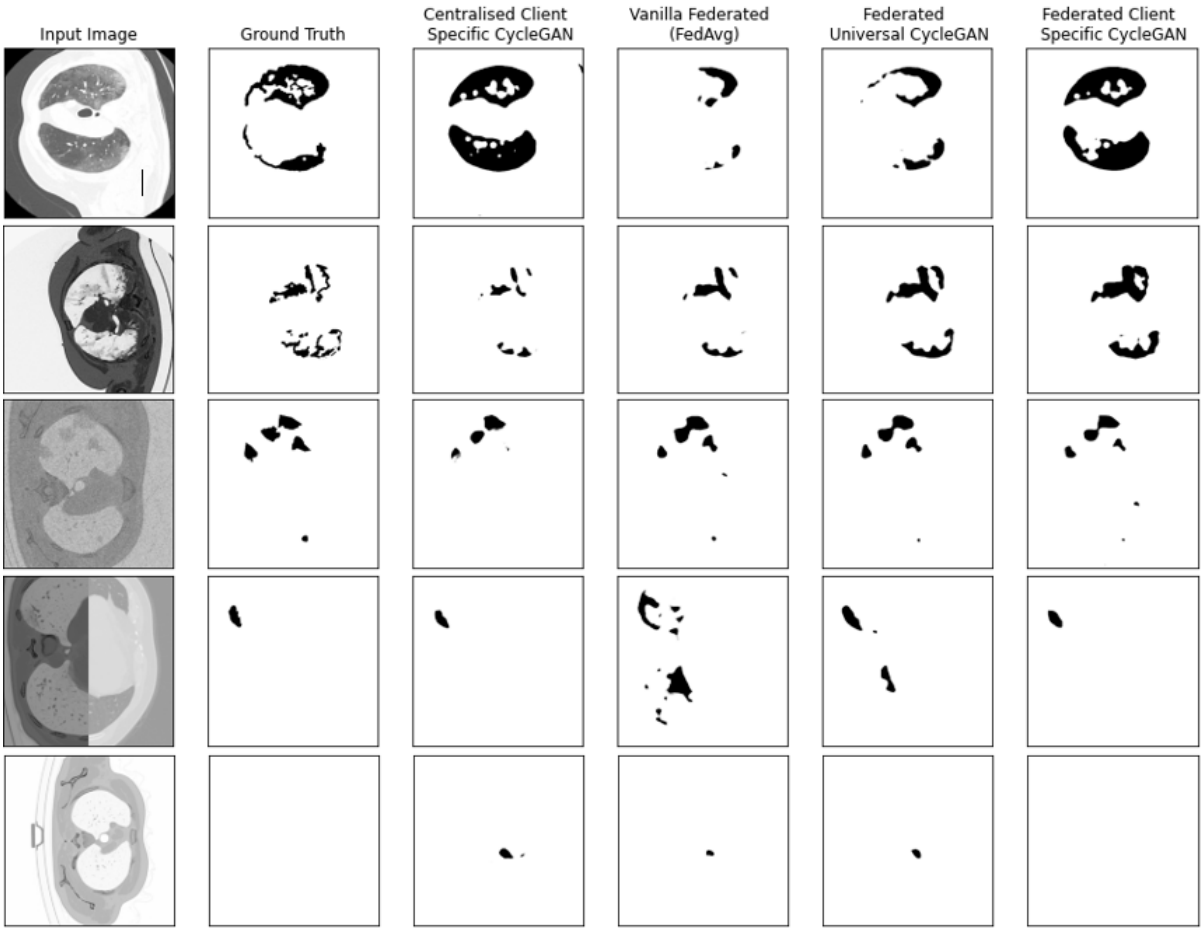

Figure 13: Sample Segmentations of Semi-Synthetic Dataset: 5 Clients