Physics in Medicine & Biology

IPEM Institute of Physics and Engineering in Medicine

OPEN ACCESS

RECEIVED
16 October 2022

REVISED
8 June 2023

ACCEPTED FOR PUBLICATION
16 June 2023

PUBLISHED
31 July 2023

PAPER

Reduction of scan duration and radiation dose in cerebral CT perfusion imaging of acute stroke using a recurrent neural network

Mahdieh Dashtbani Moghari[1,*], Amirhossein Sanaat[2], Noel Young[3,4], Krystal Moore[3], Habib Zaidi[2], Andrew Evans[5,6], Roger R Fulton[6,7,8] and Andre Z Kyme[1,8]

[1] School of Biomedical Engineering, Faculty of Engineering and Information Technologies, The University of Sydney, Sydney, Australia
[2] Geneva University Hospitals, Division of Nuclear Medicine & Molecular Imaging, CH-1205 Geneva, Switzerland
[3] Department of Radiology, Westmead Hospital, Sydney, Australia
[4] Medical imaging group, School of Medicine, Western Sydney University, Sydney, Australia
[5] Department of Aged Care & Stroke, Westmead Hospital, Sydney, Australia
[6] School of Health Sciences, University of Sydney, Sydney, Australia
[7] Department of Medical Physics, Westmead Hospital, Sydney, Australia
[8] The Brain & Mind Centre, The University of Sydney, Sydney, Australia
[*] Author to whom any correspondence should be addressed.

E-mail: dashtbani.2009@gmail.com

Keywords: deep learning, radiation dose reduction, scan duration reduction, CT perfusion imaging, acute ischaemic stroke

Abstract

Objective. Cerebral CT perfusion (CTP) imaging is most commonly used to diagnose acute ischaemic stroke and support treatment decisions. Shortening CTP scan duration is desirable to reduce the accumulated radiation dose and the risk of patient head movement. In this study, we present a novel application of a stochastic adversarial video prediction approach to reduce CTP imaging acquisition time. *Approach*. A variational autoencoder and generative adversarial network (VAE-GAN) were implemented in a recurrent framework in three scenarios: to predict the last 8 (24 s), 13 (31.5 s) and 18 (39 s) image frames of the CTP acquisition from the first 25 (36 s), 20 (28.5 s) and 15 (21 s) acquired frames, respectively. The model was trained using 65 stroke cases and tested on 10 unseen cases. Predicted frames were assessed against ground-truth in terms of image quality and haemodynamic maps, bolus shape characteristics and volumetric analysis of lesions. *Main results*. In all three prediction scenarios, the mean percentage error between the area, full-width-at-half-maximum and maximum enhancement of the predicted and ground-truth bolus curve was less than $4 \pm 4\%$. The best peak signal-to-noise ratio and structural similarity of predicted haemodynamic maps was obtained for cerebral blood volume followed (in order) by cerebral blood flow, mean transit time and time to peak. For the 3 prediction scenarios, average volumetric error of the lesion was overestimated by 7%–15%, 11%–28% and 7%–22% for the infarct, penumbra and hypo-perfused regions, respectively, and the corresponding spatial agreement for these regions was 67%–76%, 76%–86% and 83%–92%. *Significance*. This study suggests that a recurrent VAE-GAN could potentially be used to predict a portion of CTP frames from truncated acquisitions, preserving the majority of clinical content in the images, and potentially reducing the scan duration and radiation dose simultaneously by 65% and 54.5%, respectively.

1. Introduction

A multimodal Computed Tomography (CT) imaging regime, including non-contrast CT (NCCT), CT perfusion (CTP) and CT angiography (CTA), is commonly used for diagnosis and determining the best treatment options for acute ischaemic stroke patients (Ledezma and Wintermark 2009, Morgan *et al* 2015, Heit and Wintermark 2016). In quantitative CTP analysis, haemodynamic parameters such as cerebral blood volume (CBV), cerebral blood flow (CBF), mean transit time (MTT) and time to peak (TTP) are derived from CTP source data for each voxel of the brain using mathematical models (Dashtbani Moghari 2022). Based on estimated perfusion maps, the voxel-wise status of the brain tissue can be determined—specifically, the extent of hypo-perfused regions including irreversibly damaged tissue (infarct core) and potentially salvageable tissue (penumbra) (Dashtbani Moghari 2022).

A CTP imaging protocol involves the rapid acquisition of successive volumetric frames, each covering all/part of the brain, over ~1–2 min after contrast agent administration. There are two important limitations associated with this protocol. Firstly, the acquisition delivers a radiation dose of 5–6 mSv to the patient (Manniesing *et al* 2015), not excessive on its own, but large when considered alongside the dose from NCCT, CTA and other potential follow-up CT scans post treatment. Therefore, reducing the radiation dose from CTP imaging without compromising the image quality or the accuracy of haemodynamic modelling is highly desirable, especially for younger adults and paediatric patients who are more likely to be impacted by the long-term stochastic effects of ionising radiation (Wolterink *et al* 2017, Moghari *et al* 2019a). The second limitation is the risk of patient head movement during the procedure, especially during the terminal phase of the scan (Hanzelka *et al* 2013, Moghari *et al* 2019b). This movement results in streaks, distortion and blurring of reconstructed images that can impact the downstream haemodynamic modelling (Popilock *et al* 2008, Yazdi and Beaulieu 2008). An obvious motion mitigation strategy is to reduce the total scan time, however this truncates the late contrast concentration measurements which are important for image-based stroke analysis (Copen *et al* 2015, Kasasbeh *et al* 2016, Moghari *et al* 2021a).

Several paradigms have been investigated to reduce the total radiation dose in CTP imaging. The most intuitive approach is to reduce the dose per frame while preserving the total number of frames and the total scan time. Many methods have been reported to denoise low-dose CT images from shorter acquisitions, including frames from CTP imaging. Related work can be classified into sinogram domain filtration methods (Wang *et al* 2005, Karimi *et al* 2016), statistical iterative reconstruction methods (Kim *et al* 2015, Lee *et al* 2019), and image post-processing techniques that include traditional filter-based methods (Mendrik *et al* 2010, 2011, Pisana *et al* 2017) and more recent deep learning-based methods (Wolterink *et al* 2017, Moghari *et al* 2019a 2021a). The first two approaches are not practical in CTP imaging of acute stroke due to limited access to the raw sinogram data acquired on commercial scanners, high computational cost, and time delays between acquisition and reconstruction. By contrast, deep learning-based post-processing techniques are very promising for the restoration of low-dose CTP frames (e.g Kadimesetty *et al* (2018), Liu and Fang (2018), Moghari *et al* (2021a)).

A second way to reduce radiation dose in CTP imaging is to collect fewer frames by increasing the frame-to-frame time interval, and then to estimate the missing (down-sampled) frames. Missing frames can be estimated using traditional or CNN-based interpolation of the image time series (Xiao *et al* 2019, Zhu *et al* 2020), typically based on the two neighbouring frames. Potentially more reliable interpolation might be achieved using additional images in the sequence, however only if these frames are not degraded by intra- or inter-frame motion.

Although both of these paradigms are effective in reducing the radiation dose, the total scan time—and thus the likelihood of motion—is unchanged. A third approach to dose reduction in CTP is to reduce the number of frames by reducing the total scan time. This has the advantage of simultaneously reducing the radiation dose and the likelihood of head motion, which is more common in late frames. In this study, we present a novel application of a stochastic video prediction (SVP) technique to demonstrate the feasibility of this approach. Despite many interesting applications of SVP in video analysis (Kumar *et al* 2019, Villegas *et al* 2019, Franceschi *et al* 2020), to date it has not been applied to CTP data. Most SVP applications are limited to 3D spatiotemporal video data (2D + time) and are not designed to handle 4D dynamic volumetric data (3D + time) as is obtained in CTP imaging. We describe and test a SVP approach in a recurrent framework (Medsker and Jain 2001) to predict the last 8 (24 s), 13 (31.5 s), or 18 (39 s) CTP image frames from a sequence of initial 25 (36 s), 20 (28.5 s) or 15 (21 s) acquired frames, respectively. Unlike the frame interpolation techniques, our approach estimates later frames based on the preceding sequence of dynamic data. It is, therefore, a genuine predictive approach. In this work we investigate the feasibility and initial validation of the approach for clinical CTP data.

2. Materials and methods

2.1. CT perfusion data and data pre-processing

The retrospective dataset comprised CTP studies from 75 consecutive patients (42 males (56%), 33 females (44%)), who showed occlusion in the right or left middle cerebral artery on CTA, presenting to Westmead Hospital, Sydney in 2018. The average age of the patients was 71 yr (SD 15 yr, range 35–92 yr). Data collection and analysis were performed in accordance with an approved human ethics protocol.

The CTP images were acquired using a dual-source, dual-energy Siemens Somatom Force CT scanner in 4i cine mode with 1 min acquisition at 70 kVp and 200 mAs. Source rotation time was 250 ms with 1120 projection views per rotation. A standard foam headrest was used to limit patient head movement during the acquisition. The CT dose index ($\mathrm{CTDI_{vol}}$) and dose-length product of the CTP scans was 159.8 mGy and 2398.0 mGy.cm, respectively. At the start of the scan, ~45 ml of non-ionic iodinated contrast agent was administered intravenously at 7 ml $\mathrm{s^{-1}}$ via a power injector with a 5 s delay. 33 brain volumes were acquired over the 1 min scan at 1.5 s intervals for the first 25 CT volumes and 3 s intervals for the last 8 volumes. To cover the full brain, each CT volume comprised 22 axial slices with 5 mm thickness. Slices were reconstructed in a 512 × 512 matrix with 0.43 mm in-plane resolution.

All studies were pre-processed by removing the background and skull from the reconstructed image slices and scaling the intensity of brain voxels linearly from −1 to 1.

2.2. Model components and architecture

We used the stochastic adversarial video prediction (SAVP) model proposed in Lee *et al* (2018). The model combines a variational autoencoder (VAE) (Kingma and Welling 2013) and GANs (Goodfellow *et al* 2014) in a recurrent framework (Sherstinsky 2020) that allows previous outputs (predictions) to be fed back to the model as inputs. In the following sub-sections, we first describe the principle of VAE and GAN models and then the specific architecture of the proposed VAE-GAN model for application to CTP analysis (Lee *et al* 2018).

2.2.1. Variational auto encoder (VAE)

A VAE consists of two connected neural networks, the encoder and decoder. The encoder takes an input data sample x and compresses (encodes) it into a more compact representation z, known as the latent variable, in a lower dimensional space called the latent space. In latent space, similar data points are more proximate, forming clusters. The decoder learns to reconstruct (decode) the latent representation back to the original data space. To avoid discontinuities in the latent space (i.e. gaps between clusters), the posterior distribution $q(z|x)$ is calculated by assigning a mean, μ, and standard deviation, σ, to each random variable in the latent space. This stochastic generation of variables introduces local variation, resulting in a smooth latent space within and around the clusters. The encoder, however, can learn very different μ and σ values for the different classes (clusters), thus introducing discontinuity between them. Ideally, different classes should be as close to each other as possible while still being distinct, allowing for smooth sampling and efficient decoding to the data space. This proximity and differentiation requirement is enforced using Kullback–Leibler divergence (D_{KL}) in the VAE loss function. D_{KL} measures the divergence between the posterior $q(z|x)$ and prior $p(z)$ distributions, and minimizing D_{KL} optimizes μ and σ by forcing these distributions to be closer. The VAE loss ($\mathcal{L}_{\mathrm{VAE}}$) is thus given by:

$$\mathcal{L}_{\mathrm{VAE}} = \mathcal{L}_{\ell} + D_{\mathrm{KL}}, \tag{1}$$

where $\mathcal{L}_{\ell}$ and D_{KL} refer to the reconstruction error and Kullback–Leibler (KL) divergence, respectively, and are defined according to:

$$\mathcal{L}_{\ell} = -\mathbb{E}_{q(z|x)}[\log p(x|z)] \tag{2}$$

$$D_{\mathrm{KL}} = D_{\mathrm{KL}}(q(z|x)||p(z)) = \sum_{z} q(z|x) \log \frac{q(z|x)}{p(z)}, \tag{3}$$

$p(x|z)$ refers to the conditional distribution of the data x given the latent variable z and represents the likelihood of observing x given z. It is sometimes referred to as the decoder of the VAE, mapping a latent variable z to an observation x in the data space. $\mathbb{E}$ refers to the expectation operator, which calculates the average value of a function over a given probability distribution. $\mathbb{E}_{q(z|x)}$ denotes the expectation with respect to $q(z|x)$, which is the posterior distribution and represents what we know about latent variable z after observing some data. In other words, $q(z|x)$ is a probability distribution that represents our updated belief or uncertainty about the value of z after observing some data. $D_{\mathrm{KL}}(q(z|x)||p(z))$ represents the KL divergence between the probability distributions $q(z|x)$ and $p(z)$. By minimizing reconstruction error $\mathcal{L}_{\ell}$ and D_{KL}, the VAE aims to learn a good approximation of the underlying data distribution and a useful representation of the input data in the latent space.

2.2.2. Generative adversarial network (GAN)

The GAN consists of a generator network, G, and discriminator network, D, in competition. The generator learns to map (decode) the latent variables z to the data space while the discriminator, which is simply a classifier, tries to distinguish real data $x \sim p(x)$ from generated data $\hat{x} \sim p(\hat{x})$ provided by G. The objective of the GAN is to determine the binary classifier that optimally distinguishes between the real and generated data and simultaneously enables G to fit the true data distribution.

The generator and discriminator are trained alternately to minimize and maximize the objective function in turn. The objective function of the GAN, $\mathcal{L}_{\mathrm{GAN}}(G, D)$, is defined as:

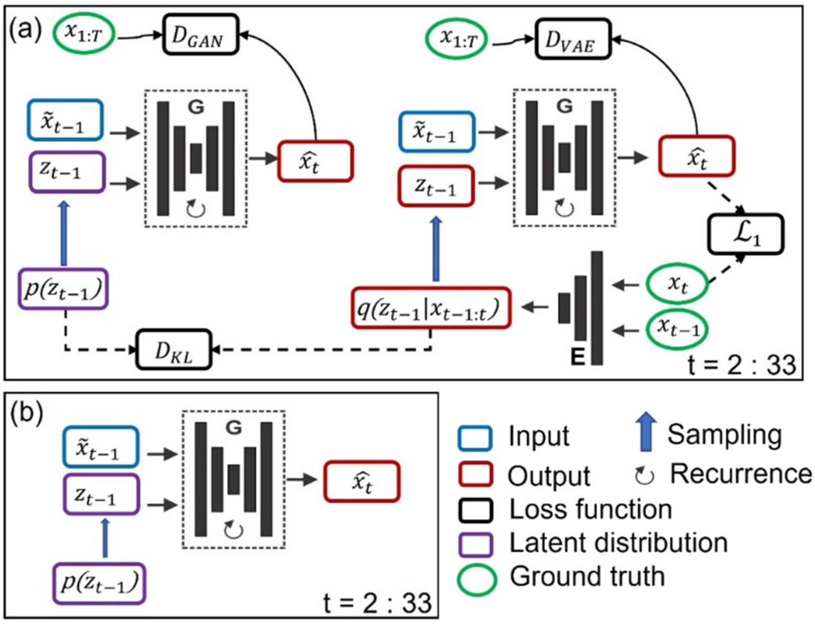

Figure 1. Block diagram of the VAE-GAN model used to predict late CTP frames $(\hat{x}_t)$ from the sequence of previous (i.e. measured) frames (x_t) shown in the training phase (a) and the testing phase (b), where $E,\ G,\ D,\ p$ and q refer to the encoder, generator, discriminator, and prior and posterior distributions, respectively, $\mathcal{L}_1$ is the $\mathcal{L}_1$ norm, and D_{KL} is the Kullback–Leibler divergence between p and q (see main text for details).

$$\mathrm{argmin}_G \mathrm{max}_D(\mathcal{L}_{\mathrm{GAN}}(G,\ D)) = \mathbb{E}_{x\char`~p(x)}[\log(D(x))] + \mathbb{E}_{\hat{x}\char`~p(\hat{x})}[\log(1 - D(\hat{x}))]. \quad (4)$$

2.2.3. Model architecture

Our VAE-GAN model was designed to predict the last 8, 13 or 18 CTP frames from the sequence of 25, 20 or 15 initial frames, respectively. The VAE component of the model is responsible for learning the underlying distribution of the input sequential data and generating a diverse set of predictions that cover a wide range of possible future frames. This is important in the context of CTP data, where tissue types can exhibit a diverse range of shapes for their time attenuation curves (TACs). However, VAE-generated predictions may lack certain details and appear blurry. To overcome this limitation, the GAN component is used to refine these predictions and make them more realistic. Thus, the combination of VAE and GAN components allows our model to generate predictions that are both diverse and realistic, sampling a range of possible futures while maintaining a high level of detail.

The model architecture is shown in figure 1. During the training phase (figure 1(a)), the recurrent generators (G) are conditioned on the previous frame $\tilde{x}_{t-1}$ and random latent code z_{t-1} to predict the next frame $\hat{x}_t$. Here the previous frame $\tilde{x}_{t-1}$ could either be a ground-truth frame x_{t-1} (as for the initial frame), or the last prediction $\hat{x}_{t-1}$. As explained in Lee *et al* (2018), the network uses a convolutional long short-term memory, implying that it can remember information derived from all earlier frames when predicting frame $\hat{x}_t$ from frame $\tilde{x}_{t-1}$. The use of a random latent code in the SAVP model can improve both the diversity and flexibility of the generated predictions, as well as the ability of the model to generalize to new and unseen situations, by combining the randomness of a latent code with the context provided by the previous frame. As shown in figure 1(a), the latent code z_{t-1} is sampled from two distributions at each time step: (1) a single posterior distribution $q(z_{t-1}|x_{t-1:t})$ estimated by an interface encoder network (E), a feed-forward network encoding two ground-truth adjacent 3D (2D space + time) frames $(x_{t-1},\ x_t)$ denoted by $x_{t-1:t}$, and (2) the prior distribution $p(z_{t-1})$, which is implemented using a standard Gaussian distribution with zero mean and unit variance. The proposed VAE-GAN model jointly optimises the VAE and GAN losses during the training. The model objective, $\mathcal{L}_{\mathrm{VAE-GANs}}$, is defined as:

$$\mathrm{argmin}_{G,E}\mathrm{max}_{D_{\mathrm{GAN}},D_{\mathrm{VAE}}}(\mathcal{L}_{\mathrm{VAE-GANs}})= \\ \lambda_1\mathcal{L}_1(G,\ E) + \lambda_{\mathrm{KL}}D_{\mathrm{KL}}(E) + \mathcal{L}_{\mathrm{GAN}}(G,\ D_{\mathrm{GAN}}) + \mathcal{L}_{\mathrm{GAN}}^{\mathrm{VAE}}(G,\ E,\ D_{\mathrm{VAE}}), \quad (5)$$

where $\mathcal{L}_1$ is the $\mathcal{L}_1$penalty between the predicted frame $\hat{x}_t$ and ground-truth frame x_t, $\mathcal{L}_{\mathrm{GAN}}$ is the objective of the GAN with discriminator D_{GAN} and latents sampled from $p(z_{t-1})$, and $\mathcal{L}_{GAN}^{VAE}$ is similar to $\mathcal{L}_{\mathrm{GAN}}$ except that it uses the latents sampled from $q(z_{t-1}|x_{t-1:t})$ and has a separate discriminator D_{VAE} (figure 1(a)). λ_1 and λ_{KL} are hyperparameters chosen by evaluating similarity metrics on the validation set during training.

During the testing phase (figure 1(b)), the generator takes in the previous frame $\tilde{x}_{t-1}$ and random latent code z_{t-1} sampled from a prior distribution $p(z_{t-1})$ to synthesise the next frame $\hat{x}_t$. The process iterates at each successive time step with the synthesised frames being fed back into the generator.

2.3. Network training and hyperparameter settings

The network was implemented in TensorFlow (Abadi *et al* 2016) and trained on a dedicated workstation with a NVIDIA GeForce RTX 2080 Ti GPU. We performed 650 000 iterations of the Adam optimizer (Kingma and Ba 2014) to train the model. The learning rate was decayed to zero linearly for the last 20 000 iterations. For the GAN models, an optimizer with $\beta_1 = 0.5$, $\beta_2 = 0.999$ and a learning rate of 0.0002 was used. The estimation error term λ_1 was set to 100 since this resulted in the best similarity performance on the validation set. For the VAE model an optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ and a learning rate of 0.001 was used.

Our data preparation step involved converting the 4D CTP data (3D space + time) into 3D (2D space + time) by stacking 2D slices from the same brain region captured at different time points. The network was trained using 65 CTP studies (47 190 2D slices) and tested using 10 randomly selected studies. The training dataset was augmented to a virtual size of 195 studies by creating 2 augmented studies for each original study using randomly applied rotation and scaling.

2.4. Model validation and performance assessment

The model described in the previous sections allows us to predict frames $\hat{x}_t,\ t \in \{t_{\text{cut}}, \ldots, 33\}$ given measured frames $x_t,\ t \in \{1, \ldots, t_{\text{cut}} - 1\}$. We tested the performance of this model for three increasingly challenging prediction tasks, corresponding to $t_{\text{cut}} = 26$, $t_{\text{cut}} = 21$ and $t_{\text{cut}} = 16$, i.e. predicting 8, 13 and 18 frames, respectively.

The most challenging prediction task involved predicting the last 18 frames from the first 15 frames. The first 15 frames often capture the wash-in of the contrast agent in the TAC, and we chose this task to evaluate how well the model predicts the entire wash-out passage, given the wash-in passage of the contrast agent. In the other two prediction tasks, some part of the wash-out passage of the contrast agent was given to the model, making these tasks less challenging than the first task.

2.4.1. Image quality metrics

The image quality of slices in predicted and ground-truth frames was compared using three metrics (Yu *et al* 2019): peak signal-to-noise ratio (PNSR), root mean squared error (RMSE) and structural similarity index (SSIM). PNSR is defined as:

$$\text{PSNR}(\hat{x}_t, x_t) = 10 \log\left(\frac{NR^2}{\|\hat{x}_t - x_t\|_2^2}\right), \tag{6}$$

where R and N denote the maximum dynamic range of the image and the total number of voxels in the brain region, respectively. PSNR is a relative image quality estimate usually expressed in decibels, higher PNSR can be indicative of better quality of the predicted set. RMSE is defined as:

$$\text{RMSE}(\hat{x}_t\ ,\ x_t) = \sqrt{\frac{\sum_N (\hat{x}_t - x_t)^2}{N}}\ , \tag{7}$$

and quantifies the discrepancy per voxel between the ground-truth and predicted images. Lower RMSE (close to zero) can indicate higher quality of estimated images. SSIM is defined as:

$$\text{SSIM}(\hat{x}_t,\ x_t) = \frac{(2\mu_{x_t}\mu_{\hat{x}_t} + c_1)(2\sigma_{x_t\hat{x}_t} + c_2)}{({\mu_{x_t}}^2 + {\mu_{\hat{x}_t}}^2 + c_1)({\sigma_{x_t}}^2 + {\sigma_{\hat{x}_t}}^2 + c_2)}, \tag{8}$$

where $\mu_{\hat{x}_t}$, μ_{x_t} and $\sigma_{\hat{x}_t}$, σ_{x_t} are the mean and variance, respectively, of $\hat{x}_t$ and x_t, $\sigma_{x_t\hat{x}_t}$ is the covariance of $\hat{x}_t$ and x_t, $c_1 = (0.01R)^2$ and $c_2 = (0.03R)^2$ (Wang *et al* 2003). Since both PSNR and RMSE are based on the mean-squared-error (MSE) between the ground-truth and predicted set, they are susceptible to bias from over-smoothing. SSIM is a useful complement to PSNR and RMSE and measures the perceived alteration in structural information between two images. SSIM ranges from 0 to 1, with higher SSIM indicating greater similarity between the two images. The metrics were calculated for each test study individually and then averaged across all 10 test studies. The same quantitative metrics were used to evaluate the associated perfusion maps derived from the CTP images.

2.4.2. Bolus shape analysis

Since the shape of the contrast bolus affects the evaluation of the tissue status, we compared the characteristics of the contrast bolus passage determined from the predicted and ground-truth images to aid assessment of the approach. The venous output function (VOF) was used to represent the bolus shape for each patient in the test group (Kasasbeh *et al* 2016). VOF is a straightforward signal to determine due to its large dimension and is less susceptible to the partial volume effect compared to the arterial input function (AIF).

The VOF was localised semi-automatically by searching for a voxel with the highest area under the time-attenuation curve (TAC) within manually defined square regions of interest (ROI) placed on the straight or sagittal sinus. The same ROIs were placed on the corresponding predicted and ground-truth images. We fitted the first-pass bolus in the VOF graph to a gamma-variate curve (Bennink *et al* 2015, Kasasbeh *et al* 2016) to obtain a robust estimate of the area under the curve (AUC), bolus peak height ($C_{\max}$), and VOF width defined as the full-width-at-half-maximum (FWHM) of the fitted gamma-variate curve. We also compared the average VOF obtained from predicted and ground-truth images for patients in the test group. As bolus arrival time was different for each patient, all VOFs were aligned to their time-to-peak before averaging (Bennink *et al* 2015).

2.4.3. Infarct and penumbra size and location

Since the size and extent of the infarct and penumbra is vital for stroke physicians to determine the best treatment options for patients, we also compared the spatial agreement of these regions derived from the ground-truth and predicted haemodynamic maps.

The sizes of the infarct and penumbra were computed by thresholding the perfusion parameters in the predicted and ground-truth haemodynamic maps using the thresholds described in (Yu *et al* 2016): the penumbral region was defined by a delay ⩾ 3 s relative to the normal hemisphere, and infarct core was defined as the sub-region of the penumbra with rCBF ⩽ 30%. The delay in the pathological hemisphere was expressed as the difference between the TTP values of each voxel in the ipsilateral hemisphere from the mean TTP of the contralateral hemisphere. rCBF denotes the percentage ratio of the CBF in the ipsilateral hemisphere to the mean CBF in the contralateral hemisphere. Infarct and penumbra volumes were calculated from the predicted and ground-truth haemodynamic maps and the average lesion size estimation error, A, and average relative lesion size estimation error, A_{rel}, were computed as (Moghari *et al* 2021a):

$$A = \frac{\sum_{i=1}^{n}(A_{\mathrm{Test}i} - A_{\mathrm{GT}i})}{n} \tag{9a}$$

$$A_{\mathrm{rel}} = \frac{\sum_{i=1}^{n}\left(\frac{A_{\mathrm{Test}i} - A_{\mathrm{GT}i}}{A_{\mathrm{GT}i}}\right)}{n}, \tag{9b}$$

where $A_{\mathrm{GT}i}$and $A_{\mathrm{Test}i}$ represent the lesion size in the ground-truth and test (predicted) images for test case i, respectively. The variable i represents the individual test case number, while n denotes the total number of test cases. For each test case, we calculated the error and relative error, and then reported the mean and standard deviation of these two metrics across all test cases.

Spatial agreement for the lesion was quantified using the dice coefficient ($F1$-score), defined as:

$$\text{Dice coefficient} = 2\left[\left(\frac{1}{\text{Precision}}\right) + \left(\frac{1}{\text{Sensitivity}}\right)\right]^{-1}, \tag{10}$$

where precision and sensitivity were defined as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{11}$$

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}, \tag{12}$$

where TP denotes true positive voxels (correctly classified infarct/penumbra tissue), FP denotes false positive voxels (healthy tissue misclassified as infarct/penumbra) and FN denotes false negative voxels (infarct/penumbra tissue misclassified as healthy) (figure 2). The highest possible value of the dice coefficient is 1, which indicates 100% spatial agreement between lesion estimates in the predicted and ground-truth data.

3. Results

3.1. Analysis of CTP images

Figure 3 shows the PSNR, RMSE and SSIM image quality results for the predicted CTP frames. These data indicate that image quality degraded monotonically from early predicted frames to late frames for all metrics and regardless of how many frames were being predicted. The ranges of average PNSR, RMSE, and SSIM were 47.48–36.50 dB, 0.004–0.015, and 0.997–0.937, respectively. The degradation was approximately linear with frame number for RMSE and SSIM. Moreover, for all three metrics the standard deviation increased over these

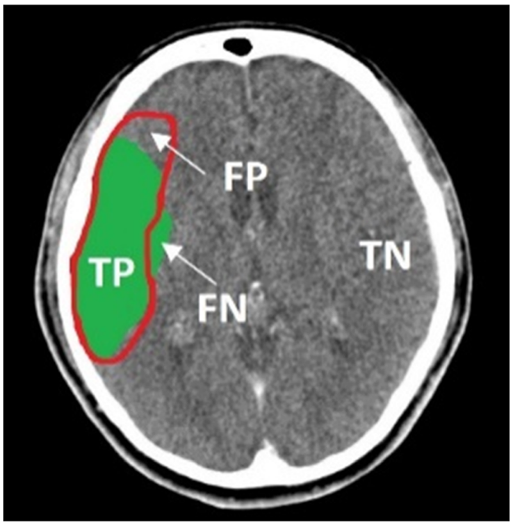

Figure 2. A typical lesion summary map. The red border indicates the lesion estimate obtained using the proposed VAE-GAN maps and the green region indicates the lesion estimate obtained using the ground-truth haemodynamic maps. The false positive (FP, healthy tissue misclassified as abnormal), false negative (FN, abnormally perfused tissue misclassified as healthy), true positive (TP, correctly classified abnormally perfused tissue) and true negative (TN, correctly classified normally perfused tissue) regions are shown for this example.

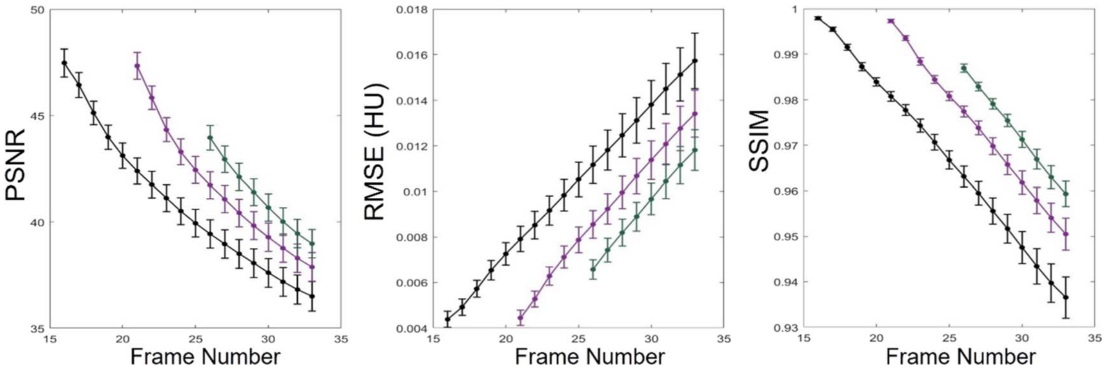

Figure 3. Average PSNR, RMSE, and SSIM (mean ± SD) of the predicted CTP frames computed for 10 test studies. In each plot the green curve shows predicted CTP frames 26–33 estimated from frames 1–25, the purple curve represents predicted CTP frames 21–33 estimated from frames 1–20, and the black curve shows the metric for predicted CTP frames 16–33 estimated from frames 1–15.

frames, suggesting that performance of the method becomes more variable the more distant the predicted frames are from measured data. Figure 4 shows some representative examples of how the image quality degradation manifests, including increased blurring of late predicted frames.

3.2. Analysis of bolus shape

Table 1 shows the results of bolus shape analysis for the three cases of predicting 8, 13 and 18 CTP frames, respectively. The lowest percentage difference between the VOF derived from predicted versus ground-truth images was observed for AUC followed (in order) by $C_{\max}$ and FWHM. Figure 5 shows the bolus shape analysis curves for the most challenging scenario of predicting the last 18 frames (typically, the entire downslope of the TAC) given the initial 15 frames.

3.3. Haemodynamics and lesion analysis

Figure 6 shows the predicted and ground-truth haemodynamic maps for CBV, CBF, MTT, and TTP. Differences are difficult to discern visually for CBV and CBF, whereas for MTT and TTP it is clear that the predicted maps tended to overestimate.

This was confirmed by the quantitative comparison of haemodynamic parameters, summarised in tables 2 and 3. The best agreement in image quality (table 2) was observed for CBV followed (in order) by CBF, MTT and TTP. Evaluation of the spatial agreement of the lesion volume in the ground-truth and predicted images (table 3)

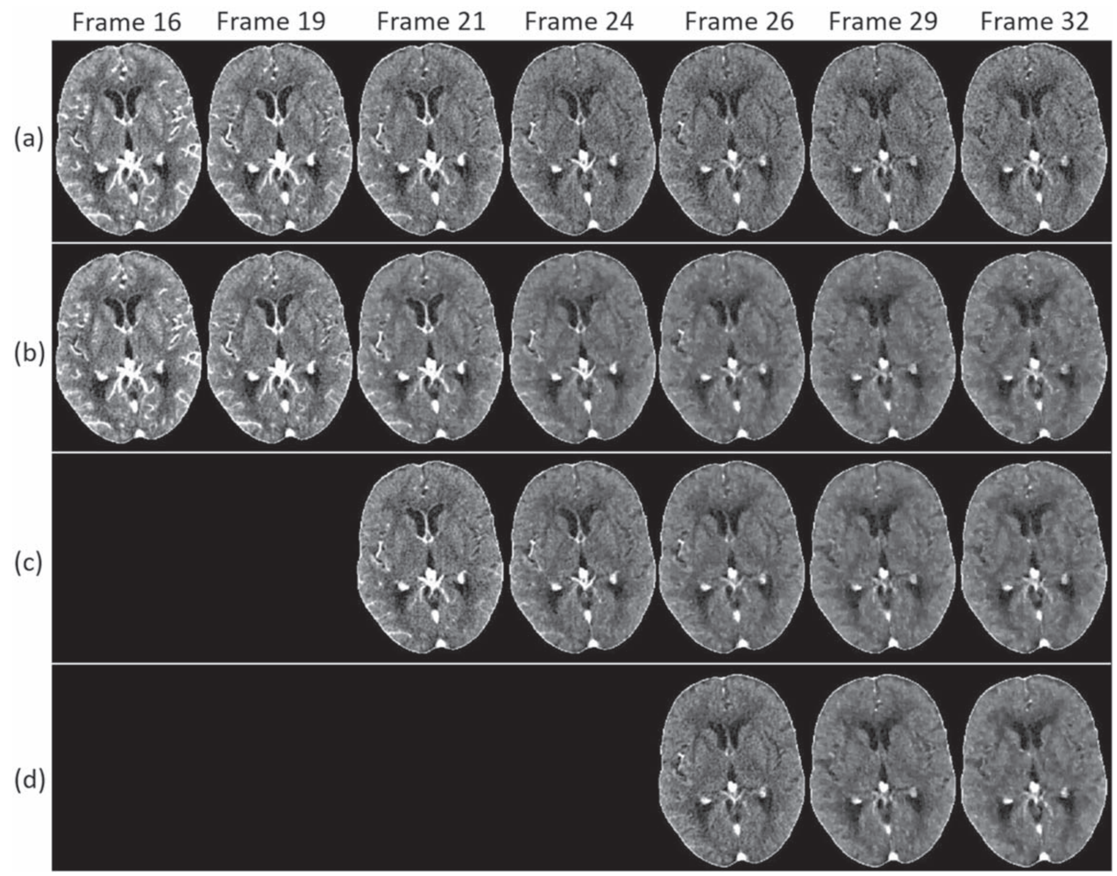

Figure 4. Examples of image quality degradation from early predicted frames to late frames. Row (a) shows ground-truth CTP frames. Rows (b)–(d) show the corresponding predicted frames estimated from frames 1–15 (b), frames 1–20 (c), and frames 1–25 (d). The pixel values are normalised between 0 (black) and 1 (white).

Table 1. Comparison of the fitted VOF characteristics (AUC, FWHM and $C_{\max}$) in the ground-truth and predicted CTP images for 3 cases: predicting frames 26–33 using frames 1–25, predicting frames 21–33 using frames 1–20, and predicting frames 16–33 using frames 1–15. Values shown represent the mean percentage error ±1 standard deviation and were computed over 10 test cases.

Predicted frames	AUC	FWHM	$C_{\max}$
26–33	1.12 ± 0.76	2.65 ± 1.94	1.77 ± 0.37
21–33	1.18 ± 0.78	2.84 ± 2.01	1.88 ± 0.43
16–33	1.70 ± 1.36	3.74 ± 3.25	2.70 ± 2.10

showed the average dice coefficient of the infarct, penumbra, and hypo-perfused region was between 67%–76%, 76%–86% and 83%–92%, respectively. The total hypo-perfused region showed greater average precision, sensitivity and dice coefficient compared to either infarct or penumbra. The lesion size error metrics (A and A_{rel} in table 3) indicated an overestimation of the average lesion volume between 7%–15% for the infarct, 11%–28% for the penumbra, and 7%–22% for the hypo-perfused regions. In all metrics, the best values were obtained when predicting the least number of frames, and the poorest values obtained when predicting the most.

4. Discussion

In this study we conditioned a SAVP approach using sequences of the first 25 (36 s), 20 (28.5 s), and 15 (21 s) reconstructed frames of a CTP study to predict the last 8 (24 s), 13 (31.5 s), and 18 (39 s) frames in order to reduce both the scan duration and the radiation dose. Feasibility of the method was assessed based on the image quality of the CTP images and haemodynamic maps, and bolus shape and volumetric lesion characterisation.

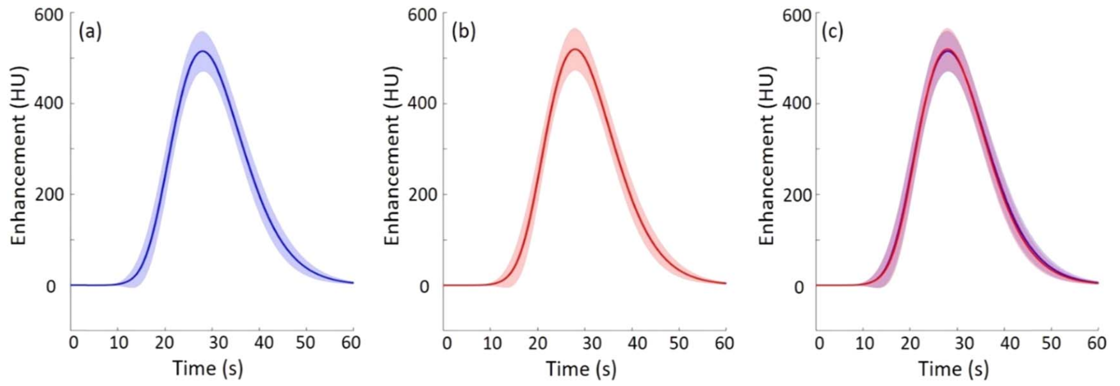

Figure 5. The mean fitted venous output function (VOF) for the case of predicting frames 16–33 using frames 1–15, computed for 10 test cases: (a) ground-truth, (b) predicted, and (c) overlay of (a) and (b). The shaded regions indicate ±1 SD. The peak enhancement of all VOFs were aligned prior to averaging.

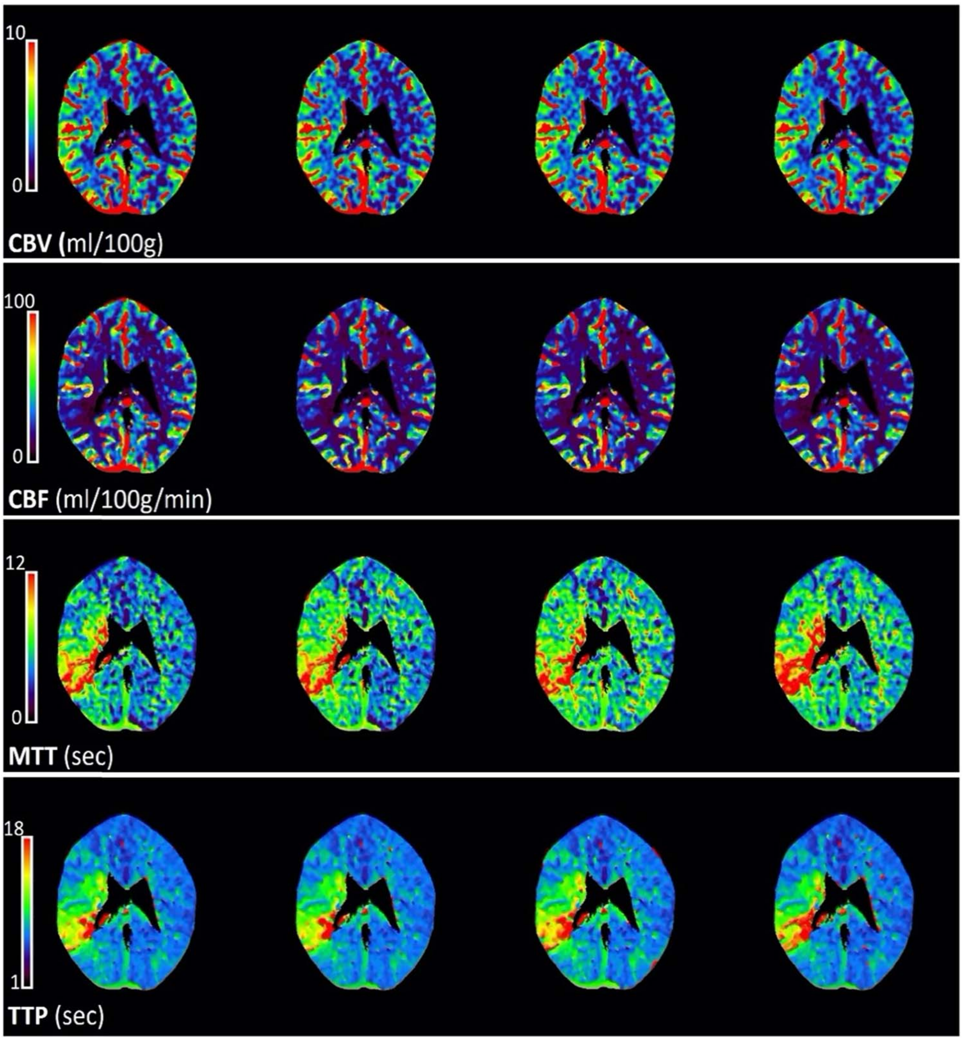

Figure 6. Ground-truth hemodynamic maps (first column) and the corresponding predicted maps estimated from frames 1–25 (second column), frames 1–20 (third column), and frames 1–15 (fourth column) using the VAE-GAN method. Rows from top to bottom show CBV, CBF, MTT and TTP maps.

The tendency of the image quality of predicted frames to degrade with frame number (figure 3), with greater error for later frames, is expected from the recurrent model since the VAE-GAN estimates each frame based on a sequence of initial ground-truth frames and previous predicted frames. The cumulative error reduces when the model makes predictions based on a higher number of initial ground-truth frames.

Table 2. PSNR, RMSE and SSIM (mean ± SD) of haemodynamic maps computed for predicted frames 26–33, 21–33, and 16–33, respectively, averaged across 10 test studies. The units of the RMSE are ml/100 g and ml/100 g min^{-1} for CBV and CBF, respectively, and seconds for MTT and TTP.

Perfusion map	Predicted frames	PSNR	RMSE	SSIM
CBV	26–33	38.24 ± 2.04	1.44 ± 0.38	0.98 ± 0.01
	21–33	36.39 ± 3.07	1.69 ± 0.42	0.98 ± 0.02
	16–33	33.42 ± 7.47	2.06 ± 0.78	0.97 ± 0.03
CBF	26–33	32.86 ± 2.08	2.74 ± 0.27	0.97 ± 0.02
	21–33	30.31 ± 2.77	2.88 ± 0.32	0.96 ± 0.02
	16–33	26.97 ± 5.35	3.10 ± 0.50	0.94 ± 0.03
MTT	26–33	25.63 ± 2.73	3.22 ± 0.36	0.91 ± 0.03
	21–33	24.10 ± 2.76	3.57 ± 0.35	0.89 ± 0.02
	16–33	21.06 ± 3.47	3.96 ± 0.45	0.87 ± 0.03
TTP	26–33	24.46 ± 2.65	3.48 ± 0.42	0.88 ± 0.04
	21–33	22.35 ± 3.87	3.86 ± 0.88	0.86 ± 0.04
	16–33	18.14 ± 4.65	4.45 ± 0.72	0.83 ± 0.05

Table 3. Lesion size characterisation (mean ± SD) computed for predicting frames 26–33, 21–33, and 16–33 from the initial 25, 20, and 15 frames, respectively, averaged across the 10 test studies.

Lesion	Predicted frames	Precision	Sensitivity	Dice	A(ml)[a]	A_{rel}[b]
Infarct	26–33	0.76 ± 0.08	0.77 ± 0.03	0.76 ± 0.06	4.00 ± 6.17	0.07 ± 0.06
	21–33	0.69 ± 0.13	0.78 ± 0.02	0.73 ± 0.08	7.02 ± 6.51	0.11 ± 0.11
	16–33	0.62 ± 0.21	0.72 ± 0.26	0.67 ± 0.23	7.58 ± 7.92	0.15 ± 0.14
Penumbra	26–33	0.77 ± 0.19	0.97 ± 0.10	0.86 ± 0.14	10.02 ± 9.32	0.11 ± 0.13
	21–33	0.75 ± 0.18	0.95 ± 0.0.09	0.84 ± 0.14	15.67 ± 13.71	0.15 ± 0.13
	16–33	0.68 ± 0.12	0.86 ± 0.07	0.76 ± 0.09	20.78 ± 12.89	0.28 ± 0.18
Hypo-perfused	26–33	0.87 ± 0.05	0.97 ± 0.03	0.92 ± 0.04	13.25 ± 4.72	0.07 ± 0.06
	21–33	0.81 ± 0.11	0.96 ± 0.03	0.88 ± 0.07	19.94 ± 11.42	0.12 ± 0.09
	16–33	0.76 ± 0.08	0.92 ± 0.09	0.83 ± 0.07	28.36 ± 14.11	0.22 ± 0.11

[a] See equation (9a).

[b] See equation (9b).

Similarly, the haemodynamics results (table 2) showed higher quality for perfusion maps computed from models using a higher number of initial ground-truth frames. The results also indicated that CBV is always the easiest parameter to predict reliably into the future, and TTP is the most difficult. A likely reason for this is that CBV is calculated from the AUC of the impulse response function, which makes it less susceptible to the TAC noise. However, TTP is estimated based on the time to the maximum (a single value) of the TAC, which is much more susceptible to noise.

Predicting a portion of CTP frames from truncated acquisitions resulted in less than 4 ± 4% difference on bolus shape metrics (average AUC, FWHM, and $C_{\max}$) compared to using all frames (table 1). However, the most important comparison from a clinical perspective (lesion analysis, table 3) indicated systematic overestimation of the infarct and penumbra in the predicted images, which reduced for models using a higher number of initial ground-truth frames. Since the infarct and penumbra volumes are calculated by thresholding the perfusion parameters, optimising the thresholds for the predicted haemodynamic maps could potentially reduce the overestimation. However, it was beyond the scope of this study to determine if such optimised thresholds exist, or to determine if the observed lesion overestimations using standard thresholds were clinically significant. Both the infarct and penumbra have delayed and very noisy TACs (figures 7(c), (d)), with the temporal changes in voxel values as low as 10–20 HU. By contrast, the TACs for regions of healthy brain (figures 7(a), (b)) exhibit a well-defined shape and temporal enhancement from the passage of contrast agent. Therefore, the prediction and differentiation of the voxels in the infarct and penumbra regions is a more challenging task compared to the healthy regions. Our results showed the lesion characterisation of hypo-perfused brain (i.e. combined infarct and penumbra) was generally better than either infarct or penumbra alone. This indicates that although our method results in a loss of accuracy in delineating these two regions, it may remain quite robust in identifying regions of reduced perfusion.

There are several potential advantages of using the proposed VAE-GAN approach in CTP imaging. Firstly, it can be used to reduce the scan duration, thereby reducing the likelihood of patient head movement during the terminal phase of the scan (Moghari et al 2021b) in addition to the radiation dose. For example, prediction of the last 18 (39 s) CTP frames from the first 15 (21 s) frames reduces the scan duration and radiation dose by around

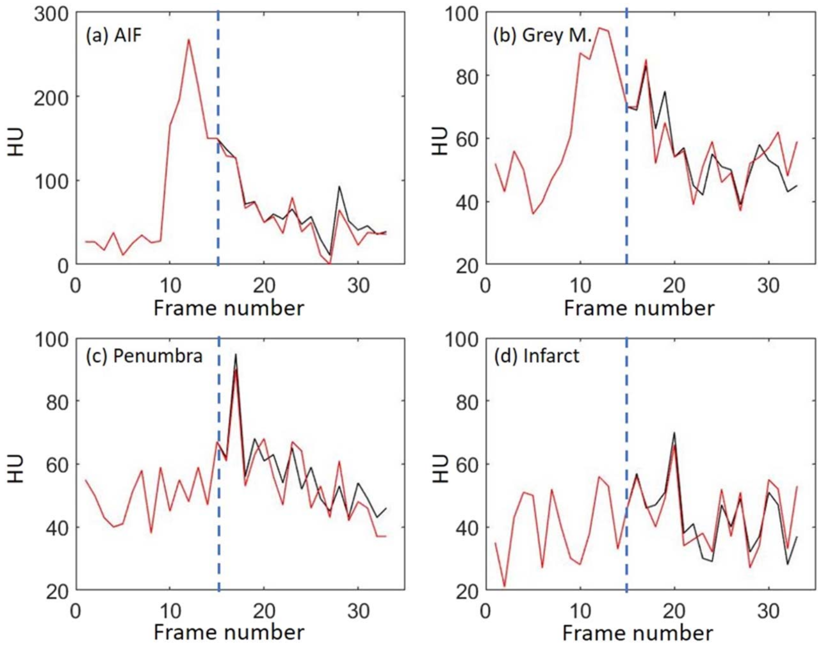

Figure 7. Examples of the time-attenuation curves (TAC) of a representative voxel in a region including (a) the arterial input function (AIF), (b) normal grey matter, (c) penumbra, and (d) infarct. In each plot, the red and black curves represent the ground-truth TAC and the TAC estimated from frames 1–15 using the VAE-GAN. Left of the vertical dashed line shows the model input.

62% and 55%, respectively. Secondly, the model was fully automated and able to predict the late frames of the whole brain volume in around 30 s in the testing phase. This is in accordance with the requirements for a practical approach in time-critical acute ischaemic stroke management. A further potential application of the method is to replace motion-corrupted frames with higher quality predicted frames in a standard CTP protocol. This will be investigated in future work.

Although our results suggest that the proposed VAE-GAN is promising as a potential practical method to reduce scan duration in CTP imaging, there are some important limitations which should temper the conclusions outlined above. Firstly, the size and diversity of the training set could affect the model performance. Our training set was relatively small (65 studies, tripled using data augmentation methods) and performance should therefore be tested for a much larger cohort of CTP studies in which the lesion size distribution is greater. Secondly, the clinical content of the predicted images must be further assessed to determine how treatment decisions would be impacted by reducing scan time. Thirdly, the generalisability of the SAVP approach in the CTP application should be evaluated beyond the single scanner and single protocol tested in this study.

5. Conclusion

In this study, we introduced and assessed a novel application of a deep learning approach to predict late CTP image frames from the early frames. The method has important potential implications for reducing the radiation dose and simultaneously reducing the probability of patient head movement in the terminal phase of the scan. Further clinical evaluation is needed to fully assess the utility of the method in practice, determining if the approach can match the clinical outcomes of analyses based on standard CTP protocols, and assessing the generalisability of the method across a more expansive training/testing set of individuals and scanners.

Acknowledgments

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Data availability statement

The data cannot be made publicly available upon publication due to legal restrictions preventing unrestricted public distribution. The data that support the findings of this study are available upon reasonable request from the authors.

Conflicts of interest

The authors declare no conflict of interest.

ORCID iDs

Habib Zaidi https://orcid.org/0000-0001-7559-5297
Roger R Fulton https://orcid.org/0000-0003-2536-2190
Andre Z Kyme https://orcid.org/0000-0003-3297-5390

References

Abadi M *et al* 2016 Tensorflow: a system for large-scale machine learning *12th {USENIX} Symp. on Operating Systems Design and Implementation ({OSDI} 16)*

Bennink E *et al* 2015 Influence of thin slice reconstruction on CT brain perfusion analysis *PLoS One* **10** e0137766

Copen W *et al* 2015 Exposing hidden truncation-related errors in acute stroke perfusion imaging *Am. J. Neuroradiol.* **36** 638–45

Dashtbani Moghari M 2022 Motion and radiation dose reduction in quantitative CT perfusion imaging of acute stroke *Phd Thesis* The University of Sydney

Franceschi J-Y *et al* 2020 Stochastic latent residual video prediction *Int. Conf. on Machine Learning* pp. 3233-3246PMLR

Goodfellow I *et al* 2014 Generative adversarial nets *Adv. Neural Inf. Process. Syst.* **27**

Hanzelka T *et al* 2013 Movement of the patient and the cone beam computed tomography scanner: objectives and possible solutions *Oral Surg. Oral Med. Oral Pathol. Oral Radiol.* **116** 769–73

Heit J J and Wintermark M 2016 Perfusion computed tomography for the evaluation of acute ischemic stroke: strengths and pitfalls *Stroke* **47** 1153–8

Kadimesetty V S *et al* 2018 Convolutional neural network-based robust denoising of low-dose computed tomography perfusion maps *IEEE Trans. Radiat. Plasma Med. Sci.* **3** 137–52

Karimi D *et al* 2016 A sinogram denoising algorithm for low-dose computed tomography *BMC Med. Imaging* **16** 1–14

Kasasbeh A S *et al* 2016 Optimal computed tomographic perfusion scan duration for assessment of acute stroke lesion volumes *Stroke* **47** 2966–71

Kim Y *et al* 2015 Ultra-low-dose CT of the thorax using iterative reconstruction: evaluation of image quality and radiation dose reduction *Am. J. Roentgenol.* **204** 1197–202

Kingma D P and Ba J 2014 Adam: a method for stochastic optimization arXiv:1412.6980

Kingma D P and Welling M 2013 Auto-encoding variational bayes arXiv:1312.6114

Kumar M *et al* 2019 Videoflow: a conditional flow-based model for stochastic video generation arXiv:1903.01434

Ledezma C J and Wintermark M 2009 Multimodal CT in stroke imaging: new concepts *Radiol. Clin. North Am.* **47** 109–16

Lee A X *et al* 2018 Stochastic adversarial video prediction arXiv:1804.01523

Lee N K *et al* 2019 Low-dose CT with the adaptive statistical iterative reconstruction V technique in abdominal organ injury: comparison with routine-dose CT with filtered back projection *Am. J. Roentgenol.* **213** 659–66

Liu P and Fang R 2018 SDCNet: smoothed dense-convolution network for restoring low-dose cerebral CT perfusion *2018 IEEE 15th Int. Symp. on Biomedical Imaging (ISBI 2018)* (IEEE)

Manniesing R *et al* 2015 Quantitative dose dependency analysis of whole-brain CT perfusion imaging *Radiology* **278** 190–7

Medsker L R and Jain L 2001 Recurrent neural networks *Des. Appl.* **5**

Mendrik A *et al* 2010 Noise filtering in thin-slice 4D cerebral CT perfusion scans *Medical Imaging: Image Processing* (International Society for Optics and Photonics)

Mendrik A M *et al* 2011 TIPS bilateral noise reduction in 4D CT perfusion scans produces high-quality cerebral blood flow maps *Phys. Med. Biol.* **56** 3857

Moghari M D *et al* 2019a Estimation of full-dose 4D CT perfusion images from low-dose images using conditional generative adversarial networks *2019 IEEE Nuclear Science Symp. and Medical Imaging Conf. (NSS/MIC)* (IEEE)

Moghari M D *et al* 2019b Characterization of the intel realsense D415 stereo depth camera for motion-corrected CT imaging *2019 IEEE Nuclear Science Symp. and Medical Imaging Conf. (NSS/MIC)* (IEEE)

Moghari M D *et al* 2021a Efficient radiation dose reduction in whole-brain CT perfusion imaging using a 3D GAN: performance and clinical feasibility *Phys. Med. Biol.* **66** 075008

Moghari M D *et al* 2021a Reducing scan duration and radiation dose in cerebral ct perfusion imaging using a recurrent neural network *2021 IEEE Nuclear Science Symp. and Medical Imaging Conf. (NSS/MIC)* (IEEE)

Moghari M D *et al* 2021b Head movement during cerebral CT perfusion imaging of acute ischaemic stroke: characterisation and correlation with patient baseline features *Eur. J. Radiol.* **144** 109979

Morgan C D *et al* 2015 Physiologic imaging in acute stroke: Patient selection *Interventional Neuroradiol.* **21** 499–510

Pisana F *et al* 2017 Noise reduction and functional maps image quality improvement in dynamic CT perfusion using a new k-means clustering guided bilateral filter (KMGB) *Med. Phys.* **44** 3464–82

Popilock R *et al* 2008 CT artifact recognition for the nuclear technologist *J. Nucl. Med. Technol.* **36** 79–81

Sherstinsky A 2020 Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network *Physica* D **404** 132306

Villegas R *et al* 2019 High fidelity video prediction with large stochastic recurrent neural networks arXiv:1911.01655

Wang J *et al* 2005 Sinogram noise reduction for low-dose CT by statistics-based nonlinear filters *Medical Imaging : Image Processing* (International Society for Optics and Photonics)

Wang Z, Simoncelli E P and Bovik A C 2003 Multiscale structural similarity for image quality assessment *The Thrity-Seventh Asilomar Conf. on Signals, Systems & Computers 2003* (IEEE)

Wolterink J M *et al* 2017 Generative adversarial networks for noise reduction in low-dose CT *IEEE Trans. Med. Imaging* **36** 2536–45

Xiao Y *et al* 2019 STIr-net: deep spatial-temporal image restoration net for radiation reduction in CT perfusion *Front. Neurol.* **10**

Yazdi M and Beaulieu L 2008 Artifacts in spiral x-ray CT scanners: problems and solutions *Int. J. Biol. Med. Sci.* **4** 135–9

Yu B *et al* 2019 Ea-GANs: edge-aware generative adversarial networks for cross-modality MR image synthesis *IEEE Trans. Med. Imaging*

Yu Y *et al* 2016 Defining core and penumbra in ischemic stroke: a voxel-and volume-based analysis of whole brain CT perfusion *Sci. Rep.* **6** 20932

Zhu H *et al* 2020 Temporally downsampled cerebral CT perfusion image restoration using deep residual learning *Int. J. Comput. Assist. Radiol. Surg.* **15** 193–201