A Robust Spatial-Temporal Multitask Deep Learning Pipeline to Predict CT Perfusion Parameters

Boan Zhu[1,*,†] Bohuai Wu[1,*] Elizabeth Tong[2] S.-H. Gary Chan[1]
[1]Hong Kong University of Science and Technology
[2]Marin General Hospital, USA
{bzhual, bwual}@connect.ust.hk liztong@gmail.com gchan@ust.hk

Abstract

Computed Tomography Perfusion (CTP) has been widely deployed to assess the blood flow in the clinical diagnosis of central nervous system. Despite being the gold standard in predicting CTP parameters, the current calculation approach based on mathematical modeling is prone to be defeated by irregular images due to artifacts (such as unpredictable imaging noise or motion blur), missing frames, or non-uniform framing intervals. To overcome that, we investigate how to predict CBV, CBF, TMAX, and MTT, the four perfusion parameters of the most clinical significance simultaneously using deep learning. We propose STM-DLP, a simple yet effective Spatial-Temporal Multitask Deep Learning Pipeline trained on the calculation results of regular images. STM-DLP extracts image features and timestamp information with spatial and temporal encoders, and subsequently correlates these features to predict the perfusion parameters in parallel. As the existing datasets do not contain framing time information, we curate a large fully timestamped real-world CTP dataset labeled fully by the calculation results and screened by neurologists and medical doctors. Through extensive experiments, we first confirm the design of STM-DLP by validating it with regular images, and then show that STM-DLP overcomes those irregular images frustrating calculation approach by predicting parameters with sound values, accommodating non-uniform framing intervals, and achieving robustness against image artifacts and reduced frames (supporting 40% frame reduction).

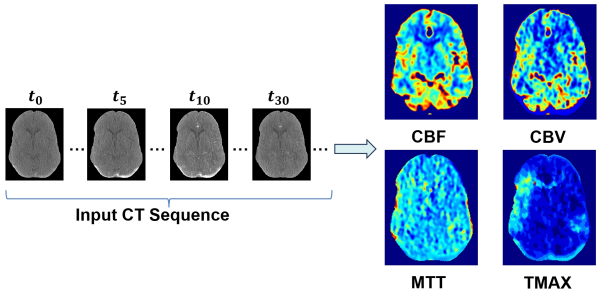


Figure 1. Illustration of the CTP perfusion parameter prediction task given a "slice," where the input comprises a (temporal) sequence of CTP images (left), and the objective is to calculate, or predict, the corresponding perfusion parameter maps for CBV, CBF, MTT and TMAX (right).

1. Introduction

Computed Tomography Perfusion (CTP) is a widely deployed imaging technique to evaluate cerebral blood flow when assessing central nervous system (CNS) conditions in the event of stroke, tumors, infections, etc. It plays a critical role in guiding decisions on medical treatment [14, 17, 19]. In CTP, the passage of an intravenously administered contrast bolus in the brain is imaged spatially and temporally by scanning different cross sections of the brain for a certain duration. This process results in a 4-dimensional (i.e., spatial-temporal) CTP dataset consisting of a stack of "slices," where a slice is a sequence of images (or frames) captured temporally at a certain location of the brain.

Many perfusion parameters can be derived given a CTP dataset [16, 20, 23]. In this paper, we focus on four parameters of the most clinical interest and significance, namely Cerebral Blood Volume (CBV, in ml/100g/min), Cerebral Blood Flow (CBF, in ml/100g), Time to Maximum (TMAX, in seconds), and Mean Transit Time (MTT, in seconds). These parameters offer crucial physiological insights into the

*Equal contribution.
†Corresponding author.

well-being of cerebral tissue. Each of these parameters is obtained by post-processing the stack of slices in the CTP dataset, resulting in values on a 2-dimensional plane for each slice. The stack of each parameter forms the so-called 3-dimensional perfusion parameter maps. Figure 1 illustrates an input CTP sequence, or "slice," on the left and the output of the four perfusion parameters of our interest on the right.

Currently the perfusion parameters are derived by calculation approach which relies on a 2-compartment kinetic models (AIF/VOF deconvolution [8, 10, 15, 18]). It has been proven to be highly accurate in computing the four perfusion parameters, and has been widely adopted clinically as the gold standard. However, such accuracy hinges upon the images being captured clearly and at regular short interval. The calculation approach is prone to be defeated by image irregularity, i.e., image artifacts arisen by, for example, random noise, patient motion blur, misalignment and missing frames, resulting in dubious or even absurd parameter values. A recent study on a common commercial calculation-based software shows a rather significant failure rate of 11% [6, 12]. For some of these failure cases, experts must make laborious and time-consuming adjustments to re-calculate the perfusion parameters. For many others, the patient has to undergo a repeated CTP scan, hence doubling the radiation risk.

For the irregular images that frustrate the calculation approach, we propose to use deep learning to predict their CTP parameters as a graceful alternative to guide medical decision under failure. Though deep learning has been studied for CT image analysis [1–5, 7, 9, 21, 22], they primarily focus on segmentation or classification of specific objects or diseases. Some works [11, 13] are related to CTP parameter prediction, however, [11] assumes that the CTP parameters are independent by using four separate single-task Convolutional Neural Network (CNN) networks. and requires framing at *fixed* intervals. [13] presents a complex GAN framework to predict CTP parameters. However, it is difficult to train and does not utilize the timestamp information in CT scans, thus vulnerable to non-uniform framing. In reality, the four CTP parameters are correlated; such correlation should be captured in order to achieve higher accuracy. Furthermore, a full CTP scan of a patient may have irregular sampling interval, skipped or unusable frames. How to effectively capture the spatial and temporal CTP features under non-uniform framing remains an open and challenging problem.

We study a multitask deep learning approach to predict CTP parameters under the general practical scenario of possibly heterogeneous framing interval. We first formulate the problem, which is to design a multitask network simultaneously predicting CBV, CBF and TMAX (note that MTT can be directly derived from CBV and CBF). We then propose STM-DLP, a simple yet effective Spatial-Temporal Multitask Deep Learning Pipeline for the problem. We train STM-DLP Using the gold standard of calculation approach as ground truth, and apply it to the irregular images of the failure cases.

As the currently available CTP datasets do not have information of framing time (i.e., assuming uniform sampling), we curate a new dataset by collecting CTP data from real patients. We generate the CTP parameters with the calculation-based RAPID software, and use the highly accurate ones screened and confirmed by medical doctors and neurologists as labels (ground truth).

By conducting extensive experiments on our real-life CTP dataset, we first validate STM-DLP design by showing its predictions to match the calculation approach for regular images (i.e., without artifacts) with low MAE values. For the irregular images of all the failure cases, STM-DLP shows strong robustness by generating parameters sound and credible to physicians. Furthermore, it requires substantially fewer frames (by 40%) to attain comparable predicted results as the calculation approach, demonstrating its applicability to the failure cases with missing frames.

2. Preliminaries

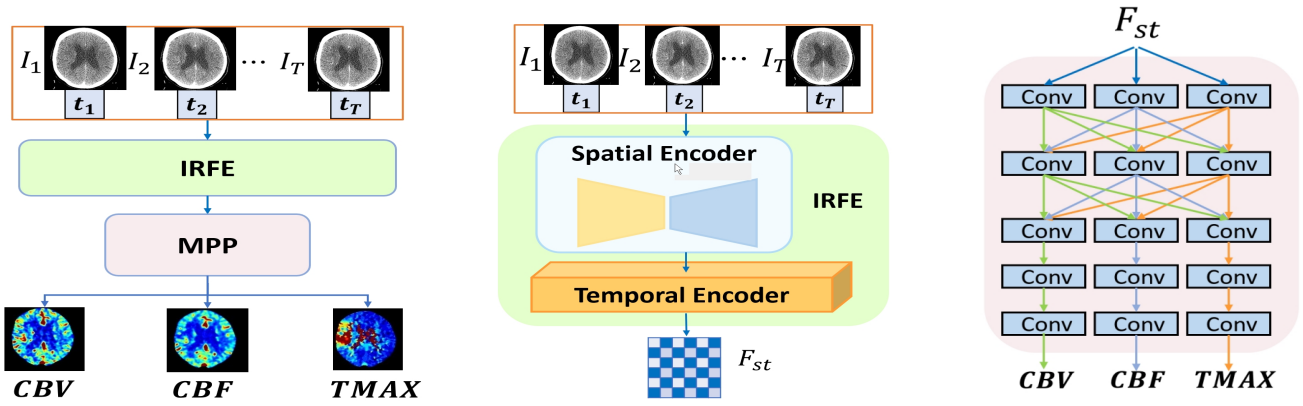
2.1. Calculation Approach

The four common CT perfusion parameters that can provide fundamental hemodynamic properties regarding the health of the cerebral tissue are:

- **Cerebral Blood Flow (CBF)**: the volume of blood passing through a given amount of brain tissue per unit time ($mL/100g/min$).
- **Cerebral Blood Volume (CBV)**: the volume occupied by intravascular blood per unit volume of brain tissue ($mL/100g$).
- **Mean Transit Time (MTT)**: the average time taken for blood to pass through a given amount of brain tissue (s).
- **Time-to-Maximum (TMAX)**: the time taken for blood to reach the maximum amplitude of the tissue response function (s).

These four parameters have been extensively studied and are widely utilized by physicians to assess the patient's CNS condition and determine the appropriate treatment.

Conventionally, the four perfusion parameters are



(a) Pipeline Overview (b) IRFE (c) MPP

Figure 2. The architecture of STM-DLP (a) consisting of IRFE (b) and MPP (c).

derived via a 2-compartmental kinetic model to formulate the brain mathematically. This model encompasses both the intravascular and the extravascular-extracellular spaces, using an artery for input and a vein for output where the model modulates the arterial input to produce the venous output. To calculate the perfusion parameters, the computational technique of deconvolution has been extensively applied in medical software such as RAPID [18]. While some derive perfusion parameters through the deconvolution of the arterial input function (AIF) from the concentration-time waveform of each voxel/pixel, others derive the parameters from the venous output function (VOF). Interested readers are referred to [10] and references therein for an in-depth explanation of the calculation process.

In the following, we overview the calculation process utilizing deconvolution of the impulse response function. The concentration of contrast medium at the voxel of interest (VOI), $c_{\text{voi}}(t)$, can be represented as a convolution expression:

$$c_{\text{voi}}(t) = (c_{\text{art}} * k_{\text{voi}})(t), \tag{1}$$

here $k_{\text{voi}}(t)$ is the flow-scaled residue function, also referred to as the impulse response function, at the voxel of interest, and $c_{\text{art}}(t)$ denotes the arterial input function (AIF). Therefore, $k_{\text{voi}}(t)$ can be derived using a deconvolution method from the measured data $c_{\text{art}}(t)$ and $c_{\text{voi}}(t)$. Then CBF is given by

$$\text{CBF} = \frac{1}{\rho_{\text{voi}}} \max(k_{\text{voi}}(t)), \tag{2}$$

where ρ_{voi} is the mean density of the volume under consideration. Also, the CBV parameter can be derived from the impulse residue function as follows:

$$\text{CBV} = \frac{1}{\rho_{\text{voi}}} \int_0^{\text{inf}} k_{\text{voi}}(t)dt, \tag{3}$$

Similarly, MTT can be determined as follows:

$$\text{MTT} = \frac{1}{\max(k_{voi}(t))} \int_0^{\text{inf}} k_{voi}(t)dt. \tag{4}$$

Then we can observe and represent the relationship among Eq.(2), Eq. (3) and Eq. (4) as $MTT = CBV/CBF$. Finally, we can calculate TMAX using the impulse response function by $\text{TMAX} = \arg\max_t k_{\text{voi}}(t)$.

According to the calculation equations, all parameters are derived from the impulse response function, underscoring the criticality of its accuracy. Nonetheless, this precision is heavily contingent upon the appropriate selection of AIF and VOF. Since the calculation model relies solely on these selections and the VOI function, errors in any element can degrade the overall results. In automated algorithms, irregular artifacts—such as sporadic noise, patient motion, and frame misalignment—often lead to suboptimal AIF/VOF selections, yielding non-diagnostic perfusion parameters. Consequently, a study reported a substantial failure rate of 11% in commonly utilized commercial software. For these failure cases, experts must either perform time-consuming manual corrections, or patients must undergo a repeat CTP scan, effectively doubling their radiation exposure.

Additionally, to ensure accuracy, this method demands a substantial number of frames, which consequently leads to long radiation exposure. In practice, this can mean at least 60 seconds of exposure with a 2-second sampling interval, often extended to 80˜120 seconds with a shorter interval to ensure accuracy.

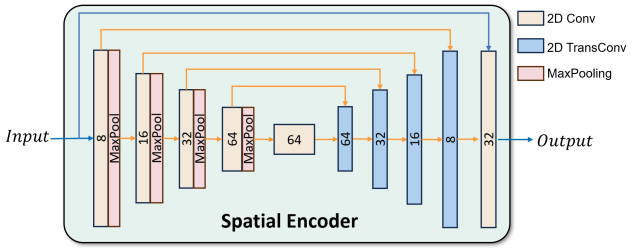


Figure 3. The architecture of the proposed spatial encoder.

Consequently, in this study, we introduce our STM-DLP technique to address the failure factors, namely, high noise levels, unusable frames resulting from irregular patient motion, and misalignment issues. Furthermore, we find that STM-DLP requires significantly fewer frames (by 40%) to achieve comparable predicted results as the calculation approach, highlighting its effectiveness in addressing failure cases where frames are missing.

3. Method

3.1. Pipeline Overview

We propose a simple yet effective Spatial-Temporal Multitask Deep Learning Pipeline (STM-DLP). As depicted in Fig. 2a, the input of STM-DLP comprises a temporal sequence of CT images and timestamp information for each image, which are processed by Impulse Response Feature Encoder (IRFE) and Multi-Parameter Predictor (MPP). The output is a set of targeted perfusion parameters, specifically CBV, CBF, and TMAX. Given the disparate range of values for each perfusion parameter, we employ visualization techniques to represent the results as color-coded images, akin to those presented in Fig. 2a.

3.2. Impulse Response Feature Encoder (IRFE)

The input 3D brain slice encompasses rich spatial information among pixels within each frame and temporal information across time intervals. Therefore, we introduce the Impulse Response Feature Encoder (IRFE) as the inaugural component, serving as an extractor to extract and integrate the spatial and temporal correlations from the sequential brain frames and generate enhanced spatial-temporal features F_{st}. By explicitly incorporating timestamp information as an input, our model can precisely capture temporal information, accommodating variable framing intervals and missing frames.

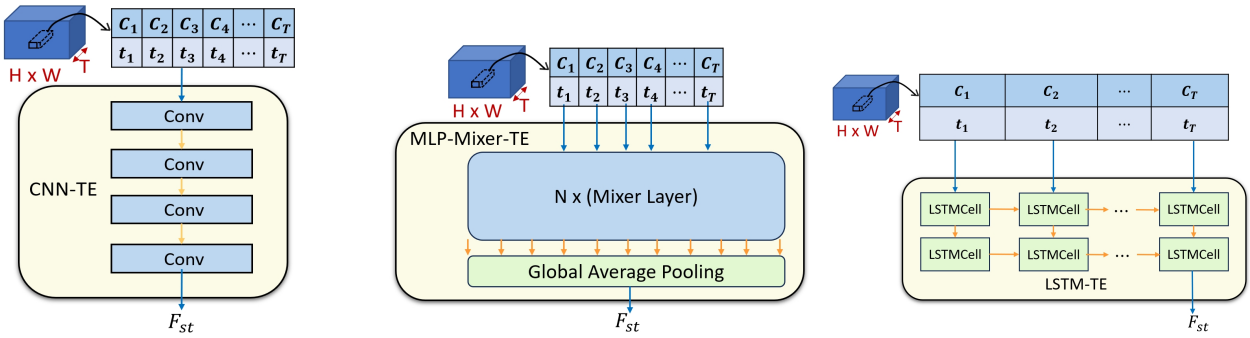
As illustrated in Fig. 2b, the IRFE comprises two sequential modules: the Spatial Encoder (Fig. 3) and the Temporal Encoder (Fig. 4). The Spatial Encoder is applied first to discern the spatial correlations among the pixels within each frame. Following this, the Temporal Encoder learns the temporal characteristics of each pixel throughout the sequence. The Spatial Encoder extracts and merges the spatial correlations, enriching each pixel's spatial features by incorporating information from neighboring pixels. These representations are then relayed to the Temporal Encoder. The Temporal Encoder focuses on extracting temporal information from the enriched spatial feature sequence. Upon completion of these steps, a comprehensive spatial-temporal feature, F_{st}, corresponding to the input sequence, is generated and forwarded to the MPP. By exploiting such sequential design, we separate the learning of spatial and temporal information to enable the encoders to perform their designated tasks independently. Furthermore, to enhance the learning for non-uniform intervals, we directly input the given time series into the temporal encoder to explicitly guide the fusion of the time-dependent features. The design of putting the spatial encoder first to learn features for each frame independently has the advantage of handling different numbers of inputs with a proper temporal encoder. In experiments, the pipeline with LSTM temporal encoder provides not only the highest performance but also the ability to predict parameters with a reduced number of frames. To further justify the design of IRFE, we conduct ablation studies in Section 4.3 to compare the performance of different pipelines with: spatial encoder first, temporal encoder first, or joint encoder, and show that our design achieves the best.

3.2.1. Spatial Encoder

Within the IRFE framework, the Spatial Encoder is structured around a UNet architecture to extract multi-scale features, as depicted in Fig. 3. The encoder segment of the UNet incorporates 2D convolution layers with 3×3 kernels, followed by max-pooling layers post each CNN block for feature map downsampling. The decoder segment of the UNet employs transposed convolution layers to amalgamate the features from the lower scale and corresponding encoder layers. The final output layer is architecturally designed to incorporate a residual connection between the input and the feature map derived from the last layer, thereby producing the output spatial feature, $F_{spatial}$. Leveraging the UNet architecture, the Spatial Encoder in IRFE can exploit and fuse spatial information among neighboring pixels across multiple scales.

3.2.2. Temporal Encoder

Acknowledging the critical role of temporal information for both the impulse response function and all associated parameters, alongside the inherent difficulty



(a) CNN Temporal Encoder (b) MLP-Mixer Temporal Encoder (c) LSTM Temporal Encoder

Figure 4. The architectures of the proposed temporal encoders.

Table 1. Quantitative comparison between UniToBrain models and STM-DLP for $D_{general}$.

Models		MAE			RMSE			R-Squared		
		CBV	CBF	TMAX	CBV	CBF	TMAX	CBV	CBF	TMAX
UniToBrain[11]		1.497	6.529	3.920	2.293	8.833	5.641	0.528	0.338	-1.864
STM-DLP	**CNN**	1.212	5.720	1.877	1.871	8.038	3.289	*0.672*	0.401	0.350
	MLP-Mixer	1.207	5.594	*1.649*	1.907	7.879	*3.085*	0.651	0.355	*0.443*
	LSTM	*1.151*	*5.429*	1.692	*1.837*	*7.860*	3.115	0.658	*0.603*	0.417

of temporal information acquisition, we implement the Temporal Encoder of IRFE to evaluate the performance capabilities of CNN, MLP-Mixer, and LSTM. The process begins with the input spatial feature $F_{spatial} \in \mathbb{R}^{T\times C\times H\times W}$ and the time series $t \in \mathbb{R}^{T}$, wherein the time series is integrated into the channel dimension for each pixel, and the spatial dimensions H, W are condensed to construct a feature vector $F \in \mathbb{R}^{HW\times T\times (C+1)}$. This approach enables the temporal encoder to decipher temporal data on a per-pixel basis. Additionally, the time values are converted into a separate channel that can be learned and integrated with other features, thereby guiding the learning process to account for non-uniform sampling intervals.

In the deployment of the Temporal Encoder using CNN, four layers of 1D convolution are applied to each pixel feature $F \in \mathbb{R}^{1\times T\times (C+1)}$ as illustrated in Fig. 4a. This convolution process occurs along the temporal dimension, facilitating the fusion and acquisition of temporal information. Consequently, this approach enables the extraction of the impulse response feature map $F_{st} \in \mathbb{R}^{H\times W\times C'}$, wherein spatial-temporal features are captured in C' dimensions for each pixel, effectively mapping the temporal features across pixels. Similarly, in the implementation utilizing MLP-Mixer shown in Fig. 4b, the temporal information learning process occurs along the temporal dimension. Here, we treat the feature $F \in \mathbb{R}^{1\times T\times (C+1)}$ associated with each pixel as T "patches", each comprising $C+1$ features, in line with the MLP-Mixer's original design that segments images into patches akin to the Vision Transformer [9], thereby facilitating feature learning. In the final stage, global average pooling is employed to compute the mean of all features across the T "patches," which serves as the output. In the LSTM-based implementation of the Temporal Encoder, we apply a two-layer LSTM to learn the features along the temporal dimension, as depicted in Fig. 4c. The recurrent neural network (RNN) mechanism inherent to LSTM allows the Temporal Encoder to manage varying numbers of input CT frames, thereby enhancing the efficiency of our STM-DLP in predicting parameters using fewer frames.

3.3. Multi-Parameter Predictor (MPP)

To simultaneously predict all perfusion parameters, our pipeline adopts a multitask approach that leverages the IRFE to derive a comprehensive spatial-temporal feature F_{st}, followed by the MPP for concurrent parameter estimation. Utilizing the shared feature F_{st} as the input, MPP employs three parallel modules, each dedicated to the estimation of a distinct parameter, as shown in Fig. 2c. To capitalize on inter-parameter correlations, MPP facilitates feature exchange among different tasks at the initial n layers (where n is a hyperparameter), enhancing prediction accuracy by integrating attributes from the diverse tasks. After feature exchange and amalgamation, each predictor au-

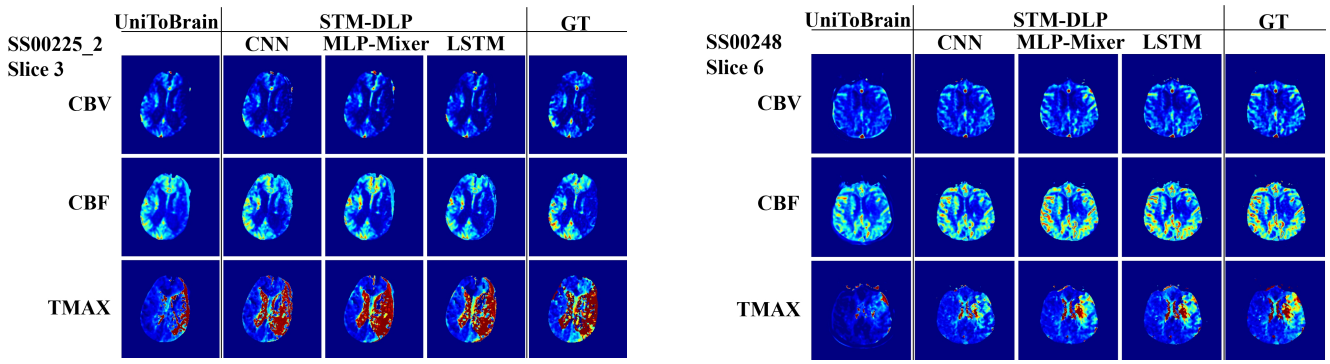


Figure 5. Qualitative comparison between UniToBrain models and STM-DLP on $D_{general}$.

tonomously estimates a single parameter to prevent cross-task interference. Our proposed framework thus concentrates on multitask learning of perfusion parameters, using the impulse response feature as a foundation. This multitask configuration allows the network to predict all CTP parameters collectively, achieving high precision for individual parameters within a unified architecture.

4. Experiments

4.1. Experimental Setup

To evaluate the effectiveness of the STM-DLP, we compile datasets from real-life scenarios encompassing various conditions. The first dataset, labeled as $D_{general}$, comprises data from 195 patients. From each patient's 4D data (a stack of brain slices), we extract a total of 2356 independent 3D slices. Each slice contains a sequence of images (frames) captured at a specific brain cross-section over time. Each patient's perfusion parameters are generated by RAPID [18], a gold-standard commercial software used in hospitals. By processing the 4D CTP data, RAPID calculates the ground-truth parameters for each slice. Doctors and neurologists carefully review and select the data to ensure the accuracy of the perfusion parameters. Since these parameters are directly used in patient treatments, their accuracy is further validated, making them our ground truth and reference. For model evaluation, the dataset is divided into training and testing subsets, containing 2,282 and 74 data instances, respectively. The division of the dataset is performed based on individual patients, ensuring that there is no overlap between the training and testing subsets, which helps prevent overfitting. Furthermore, an additional dataset, D_{fail}, consists of 68 slices from exclusive 7 patients, characterized by the failure of conventional calculation methods to determine perfusion parameters. This dataset underscores the challenging cases our model aims to address. The datasets encompass data from distinct patients, featuring non-uniform sampling intervals and varying frame counts (e.g., 46/50/55 frames for different patients), which adds to the complexity of the evaluation. To quantitatively assess the performance of STM-DLP, we utilize the MAE, RMSE and R-squared metrics.

4.2. Implementation Details

The training process consists of 200 epochs. We employ the Adam optimizer with momentum parameters set to $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and weight decay set to 0. The Charbonnier loss is used with equal weighting for each task during training. The initial learning rate is set to 5×10^{-4}. A StepLR learning rate scheduler is applied to reduce the learning rate by a factor of 0.995 every 10 epochs. All experiments are conducted on NVIDIA RTX4090 GPUs.

4.3. Ablation Study

Ablation studies are conducted on the dataset to validate the design of our pipeline, as illustrated in Table 2. As described in Section 3.2, we set up 3 models by putting the spatial encoder first, the temporal encoder first, or utilizing a joint encoder. We also provide studies on different positions in which the time series is fed. For the **Temporal First** and **Joint Encoder** models, the input positions of the time series are limited to being placed together with the input frames at the beginning. For the **Spatial First** models, we compare the performances that the time series is inputted at the temporal encoder, spatial encoder, or both. The empirical results indicate that the **Spatial First** models outperform both the **Temporal First** and **Joint Encoder** configurations. Notably, the **Spatial First** model, with the time series introduced at the temporal encoder, demonstrates superior performance. This finding corroborates the efficacy and strategic design of our pipeline.

We also conduct an ablation study on the number

Table 2. Ablation studies on IRFE design with LSTM implementation on $D_{general}$.

Models	MAE			RMSE			R-Squared		
	CBV	CBF	TMAX	CBV	CBF	TMAX	CBV	CBF	TMAX
Temporal First	1.152	5.434	1.888	1.836	7.772	3.461	0.680	0.385	0.309
Spatial First - Time in Temporal	*1.151*	*5.429*	*1.692*	1.847	7.860	*3.115*	0.658	*0.603*	*0.417*
Spatial First - Time in Spatial	1.177	5.853	1.751	1.864	8.242	3.211	0.647	0.269	0.380
Spatial First - Time in both	1.210	5.599	1.748	1.917	8.061	3.186	0.626	0.322	0.401
Joint Encoder	1.170	5.449	2.178	*1.832*	*7.496*	3.662	*0.709*	0.511	0.230
MPP with 1-layer Feature Exchange	*1.126*	5.525	1.731	*1.777*	7.910	3.133	*0.699*	0.596	0.407
MPP with 2-layer Feature Exchange	1.151	*5.429*	*1.692*	1.847	*7.860*	*3.115*	0.658	*0.603*	*0.417*
MPP with 3-layer Feature Exchange	1.153	5.542	1.721	1.831	8.013	3.147	0.669	0.544	0.408

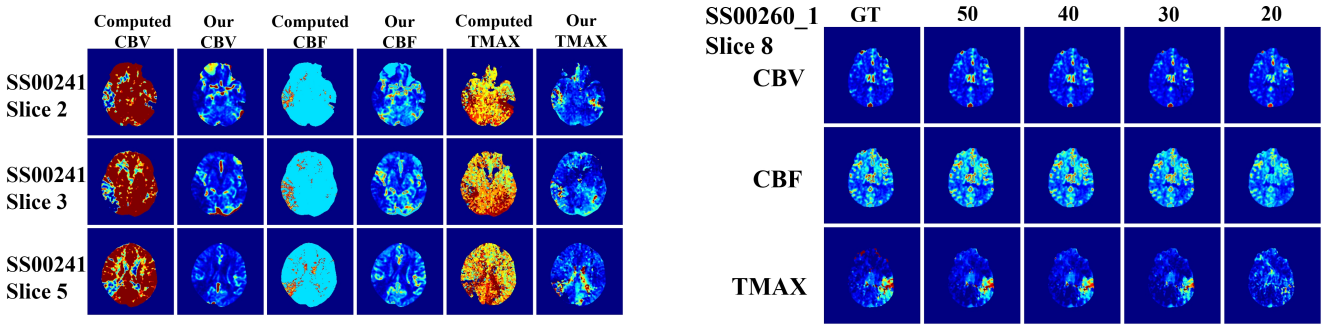


Figure 6. Qualitative comparison between the commercial software and the STM-DLP on D_{fail} (left) and reduced frames on $D_{general}$ (right).

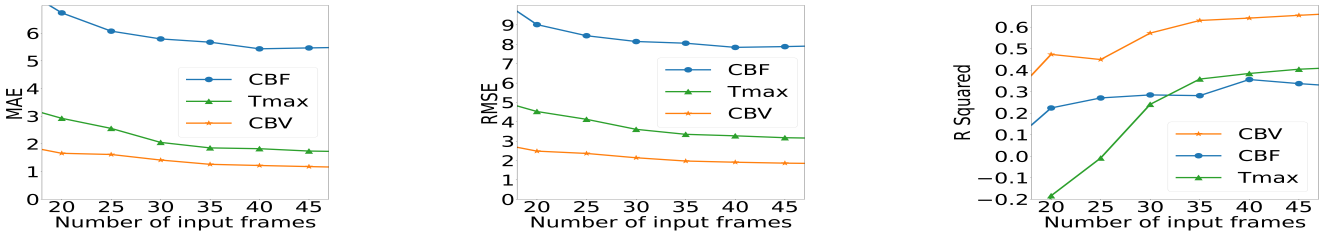
of feature exchange layers in MPP. As mentioned in Section 3.3, we set a total of 5 layers: the first layer processes the spatial and temporal feature F_{st}, the last layer generates the output, and the remaining 3 layers are used for feature exchange. We compare different models using 1, 2, or 3 layers for feature exchange. As shown in Fig. 2c, MPP with 2 layers of feature exchange achieves better performance on most metrics.

4.4. Performance Comparison

We implement the STM-DLP using the most commonly used deep learning modules (CNN, LSTM, and MLP-Mixer) to evaluate and compare their performance. For the UniToBrain model [11], we train 3 independent models for each parameter separately.

General performance: We conduct experiments on $D_{general}$ to validate the general performance of STM-DLP with our different implementations on regular cases. The quantitative comparison results are presented in Table 1. Compared with three independent UniToBrain models, all implementations of our pipeline achieve better performance significantly. The results demonstrate that the STM-DLP with LSTM outperforms the other models across CBV and CBF on most evaluation metrics. The pipeline with MLP-Mixer exhibits solid performance on TMAX. In addition, Figure 5 illustrates a qualitative comparison of the models. Perfusion parameters are visualized in color with high values represented by colors close to red, while low values are depicted with blue colors. Typically, physicians can identify abnormal regions based on the areas' color. It is evident from the figure that STM-DLP with LSTM can achieve better predictions than the other models, especially on the regions' color predictions and consistency. Moreover, MLP-Mixer-based STM-DLP has better performance on the normal regions (regions in blue), which brings the best quantitative performance on TMAX. Compared to independent models, the results demonstrate the advantages of multitask learning, which captures the correlations among parameters, thereby improving performance across all parameters. Due to its superior performance and the ability to process any number of input frames, we utilize the LSTM-based STM-DLP to test the following cases.

Robustness to failure cases: A critical bottleneck of mathematical CTP modeling is its susceptibility to imaging noise and non-uniform framing. To validate the robustness of STM-DLP, we conducted zero-shot evaluations on the D_{fail} dataset. Given the absence of explicit ground truth in corrupted data, validation inherently relies on clinical coherence. Figure 6 (left) demonstrates that while conventional calculation methods suffer from catastrophic degrada-



(a) MAE (b) RMSE (c) R Squared

Figure 7. Quantitative comparison plots of the STM-DLP model with reduced frames using dataset $D_{general}$ via (a) MAE, (b) RMSE, and (c) R Squared.

tion—yielding uniform and diagnostically meaningless regions—STM-DLP preserves robust feature extraction. By circumventing the ill-posed inverse problem of deconvolution, our model synthesizes anatomically consistent perfusion maps. Independent evaluations by neuroradiologists verified that the STM-DLP predictions align with physiological expectations, successfully mitigating extreme artifact interference and providing a credible diagnostic reference where traditional pipelines fail.

Reducing the number of input frames: Further experiments are conducted to assess the STM-DLP model's efficacy with a reduced number of frames on the $D_{general}$ dataset. This investigation aims to emulate the situation where later frames are lost during the scanning process while maintaining the original framing intervals. By using only the initial set of frames, the model can effectively handle scenarios where the later frames are missing. Figure 6 (right) provides a visual comparison of the model's performance with varying numbers of frames and Figure 7 depicts the relationship between the reduction in the number of input frames and the corresponding decline in model performance through graphs. Notably, despite a decrease in the number of input frames, the STM-DLP model retains its capability to generate reliable predictions, even with as few as 30 frames. By utilizing the initial 30 out of 50 frames, the STM-DLP model demonstrates that it can still deliver performance comparable to that of the full sequence, even when the later frames are lost. This outcome demonstrates the STM-DLP's efficiency in reducing the requisite number of frames for accurately predicting perfusion parameters, highlighting its applicability to the cases with missing frames.

5. Limitations and Future Work

While STM-DLP has demonstrated strong performance and robustness in predicting CTP parameters, our current study has certain limitations. Primarily, STM-DLP is proposed as a flexible and generalized pipeline, meaning that a wide variety of advanced deep learning architectures could theoretically be integrated into the Impulse Response Feature Encoder (IRFE) and Multi-Parameter Predictor (MPP). However, due to space and computational resource constraints, we restricted our current implementations to three widely used modules (CNN, MLP-Mixer, and LSTM) to validate the overall feasibility and efficacy of the pipeline. In future work, we plan to explore and implement a broader range of state-of-the-art architectures to further unlock the potential of the proposed framework and validate its performance on larger, multi-center cohorts.

6. Conclusion

In this paper, we propose STM-DLP, a simple yet effective spatial-temporal multitask deep learning pipeline consisting of two key modules: Impulse Response Feature Encoder (IRFE) and Multi-Parameter Predictor (MPP). IRFE, in the form of a spatial encoder followed by a temporal encoder, effectively learns and fuses spatial and temporal features in CTP images and timestamp information. The MPP module leverages the correlations among parameters and simultaneously predicts the perfusion parameters. We have conducted extensive experiments on our newly collected real-life CTP datasets. Our results show that the STM-DLP model, as compared with the conventional calculation approach, achieves similar performance for the general cases with regular images (without artifacts). It is also robust against the impact of irregular images (with artifacts) or non-uniform framing and able to generate parameters credible to physicians for all the failure cases of the calculation approach. Furthermore, it requires much fewer frames (up to 40% fewer frames) to achieve performance comparable to that of the calculation approach.

Acknowledgments

This work was supported, in part, by RGC-General Research Fund (under grant number 16201625), and Smart Traffic Fund (under grant number STF26EG01) of Hong Kong.

References

[1] S Mazdak Abulnaga and Jonathan Rubin. Ischemic stroke lesion segmentation in CT perfusion scans using pyramid pooling and focal loss. In *Proc. MICCAI workshop*, pages 352–363, Granada, Spain, 2019. Springer. 2

[2] Isah Salim Ahmad, Jingjing Dai, Yaoqin Xie, and Xiaokun Liang. Deep learning models for ct image classification: a comprehensive literature review. *Quantitative Imaging in Medicine and Surgery*, 15(1):962–1011, 2025.

[3] Kimberly Amador, Matthias Wilms, Anthony Winder, Jens Fiehler, and Nils D Forkert. Predicting treatment-specific lesion outcomes in acute ischemic stroke from 4D CT perfusion imaging using spatio-temporal convolutional neural networks. *Medical Image Analysis*, 82:102610, 2022.

[4] Kimberly Amador, Anthony Winder, Jens Fiehler, Matthias Wilms, and Nils D Forkert. Hybrid spatio-temporal transformer network for predicting ischemic stroke lesion outcomes from 4D CT perfusion imaging. In *Proc. MICCAI*, pages 644–654. Springer, 2022.

[5] Reza Azad, Amirhossein Kazerouni, Moein Heidari, Ehsan Khodapanah Aghdam, Amirali Molaei, Yiwei Jia, Abin Jose, Rijo Roy, and Dorit Merhof. Advances in medical image analysis with vision transformers: A comprehensive review. *arXiv preprint arXiv:2301.03505*, 2023. 2

[6] Saif Bushnaq, Ameer E Hassan, Adam Delora, Ali Kerro, Anita Datta, Rime Ezzeldin, Zuhair Ali, Tunmi Anwoju, Layla Nejad, Rene Silva, et al. A comparison of CT perfusion output of rapid. ai and viz. ai software in the evaluation of acute ischemic stroke. *American Journal of Neuroradiology*, 2024. 2

[7] Yixin Chen, Yajuan Gao, Lei Zhu, Wenrui Shao, Yanye Lu, Hongbin Han, and Zhaoheng Xie. PCNet: Prior category network for CT universal segmentation model. *IEEE Transactions on Medical Imaging*, 2024. 2

[8] Ezequiel de la Rosa, David Robben, Diana M Sima, Jan S Kirschke, and Bjoern Menze. Differentiable deconvolution for improved stroke perfusion analysis. In *Proc. MICCAI*, pages 593–602, Lima, Peru, 2020. Springer. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 2, 5

[10] Andreas Fieselmann, Markus Kowarschik, Arundhuti Ganguly, Joachim Hornegger, and Rebecca Fahrig. Deconvolution-based CT and MR brain perfusion measurement: theoretical model revisited and practical implementation details. *Journal of Biomedical Imaging*, 2011:1–20, 2011. 2, 3

[11] Umberto A Gava, Federico D'Agata, Enzo Tartaglione, Riccardo Renzulli, Marco Grangetto, Francesca Bertolino, Ambra Santonocito, Edwin Bennink, Giacomo Vaudano, Andrea Boghi, et al. Neural network-derived perfusion maps: a model-free approach to computed tomography perfusion in patients with acute ischemic stroke. *Frontiers in Neuroinformatics*, 17: 852105, 2023. 2, 5, 7

[12] Frans Kauw, Jeremy J Heit, Blake W Martin, Fasco van Ommen, L Jaap Kappelle, Birgitta K Velthuis, Hugo WAM de Jong, Jan W Dankbaar, and Max Wintermark. Computed tomography perfusion data for acute ischemic stroke evaluation using rapid software: pitfalls of automated postprocessing. *Journal of Computer Assisted Tomography*, 44(1):75–77, 2020. 2

[13] Wasif Khan, Kyle B See, Simon Kato, Ziqian Huang, Amy Lazarte, Kyle Douglas, Xiangyang Lou, Teng J Peng, Dhanashree Rajderkar, John Rees, et al. Physiology-informed generative multi-task network for contrast-free ct perfusion. *arXiv preprint arXiv:2505.22673*, 2025. 2

[14] Nak-Hoon Kim, Sue Young Ha, Gi-Hun Park, Jong-Hyeok Park, Dongmin Kim, Leonard Sunwoo, Min-Surk Kye, Sung Hyun Baik, Cheolkyu Jung, Wi-Sun Ryu, et al. Comparison of two automated CT perfusion software packages in patients with ischemic stroke presenting within 24 h of onset. *Frontiers in Neuroscience*, 18:1398889, 2024. 1

[15] Ke Li and Guang-Hong Chen. Statistical properties of cerebral CT perfusion imaging systems. part ii. deconvolution-based systems. *Medical physics*, 46(11): 4881–4897, 2019. 2

[16] Peirong Liu, Yueh Z Lee, Stephen R Aylward, and Marc Niethammer. Perfusion imaging: An advection diffusion approach. *IEEE Trans. Med. Imag*, 40(12): 3424–3435, 2021. 1

[17] Sibo Liu, Fushun Piao, and Kai Xu. Commentary: insights and future directions from a ct perfusion-based study on pocd risk prediction. *International Journal of Surgery*, 112(1):2000–2001, 2026. 1

[18] L Luqi and R Steigerwald. Rapid software prototyping. In *Proc. HICSS*, pages 470–479. IEEE, 1992. 2, 3, 6

[19] Alejandro Rodríguez-Vázquez, Carlos Laredo, Luis Reyes, Guillem Dolz, Antonio Doncel-Moriano, Laura Llansó, Salvatore Rudilosso, Laura Llull, Arturo Renú, Sergio Amaro, et al. Computed tomography perfusion as an early predictor of malignant cerebral infarction. *European Stroke Journal*, 10(1):172–180, 2025. 1

[20] Eve S Shalom, Amirul Khan, Sven Van Loo, and Steven P Sourbron. Current status in spatiotemporal analysis of contrast-based perfusion mri. *Magnetic Resonance in Medicine*, 91(3):1136–1148, 2024. 1

[21] Peng Yang, Yuchen Zhang, Haijun Lei, Yueyan Bian, Qi Yang, and Baiying Lei. Acute ischemic stroke onset time classification with dynamic convolution and perfusion maps fusion. In *Proc. MICCAI*, pages 558–568. Springer, 2023. 2

[22] Yi Yang, Jinjun Yang, Jiao Feng, Yi Wang, et al. Early diagnosis of acute ischemic stroke by brain computed tomography perfusion imaging combined with head and neck computed tomography angiography on deep learning algorithm. *Contrast Media & Molecular Imaging*, 2022, 2022. 2

[23] Qihao Zhang, Pascal Spincemaille, Michele Drotman, Christine Chen, Sarah Eskreis-Winkler, Weiyuan Huang, Liangdong Zhou, John Morgan, Thanh D Nguyen, Martin R Prince, et al. Quantitative transport mapping (qtm) for differentiating benign and malignant breast lesion: Comparison with traditional kinetics modeling and semi-quantitative enhancement curve characteristics. *Magnetic Resonance Imaging*, 86:86–93, 2022. 1