A General and Scalable Vision Framework for Functional Near-Infrared Spectroscopy Classification

Zenghui Wang, Jun Zhang, Yi Xia, Peng Chen, *Member, IEEE*, and Bing Wang, *Senior Member, IEEE*

***Abstract*— Functional near-infrared spectroscopy (fNIRS), a non-invasive optical technique, is widely used to monitor brain activities for disease diagnosis and brain-computer interfaces (BCIs). Deep learning-based fNIRS classification faces three major barriers: limited datasets, confusing evaluation criteria, and domain barriers. We apply more appropriate evaluation methods to three open-access datasets to solve the first two barriers. For domain barriers, we propose a general and scalable vision fNIRS framework that converts multi-channel fNIRS signals into multi-channel virtual images using the Gramian angular difference field (GADF). We use the framework to train state-of-the-art visual models from computer vision (CV) within a few minutes, and the classification performance is competitive with the latest fNIRS models. In cross-validation experiments, visual models achieve the highest average classification results of 78.68% and 73.92% on mental arithmetic and word generation tasks, respectively. Although visual models are slightly lower than the fNIRS models on unilateral finger- and foot-tapping tasks, the F1-score and kappa coefficient indicate that these differences are insignificant in subject-independent experiments. Furthermore, we study fNIRS signal representations and the classification performance of sequence-to-image methods. We hope to introduce rich achievements from the CV domain to improve fNIRS classification research.**

***Index Terms*— Functional near-infrared spectroscopy (fNIRS), brain–computer interfaces (BCIs), classification, deep learning, Gramian angular difference field.**

Manuscript received 27 January 2022; revised 19 April 2022 and 26 May 2022; accepted 29 June 2022. Date of publication 13 July 2022; date of current version 22 July 2022. This work was supported by the National Natural Science Foundation of China under Grant 61872004, Grant 62072002, and Grant 62172004. *(Corresponding author: Jun Zhang.)*

Zenghui Wang is with the School of Electrical Engineering and Automation, Anhui University, Hefei 230601, China (e-mail: scholarzhwang@163.com).

Jun Zhang and Yi Xia are with the School of Artificial Intelligence, Anhui University, Hefei 230601, China (e-mail: junzhang@ahu.edu.cn; xiayi@ahu.edu.cn).

Peng Chen is with the National Engineering Research Center for Agro-Ecological Big Data Analysis and Application, School of Internet, and Institutes of Physical Science and Information Technology, Anhui University, Hefei 230601, China (e-mail: pengchen@ustc.edu).

Bing Wang is with the School of Electrical and Information Engineering, Anhui University of Technology, Maanshan 243002, China (e-mail: wangb@ahut.edu.cn).

Digital Object Identifier 10.1109/TNSRE.2022.3190431

I. Introduction

FUNCTIONAL near-infrared spectroscopy (fNIRS) is a new non-invasive neuroimaging technique that measures brain hemodynamic responses using near-infrared light between 650 and 950 nm [1]. Neuronal activities consume oxygen carried by hemoglobin during brain tissue metabolism, leading to changes in the concentrations of oxygenated hemoglobin (HbO) and deoxygenated hemoglobin (HbR) in activated regions [2]. Coyle *et al.* [3] demonstrated for the first time that fNIRS can be utilized to develop brain-computer interfaces (BCIs) that provide non-muscular support for patients with severe motor impairment. The advantages of fNIRS, such as its safety, mobility, and low noise level, have sparked considerable interest in BCIs [2]. As a result, fNIRS classification has become the focus of research.

Statistical features, such as mean, variance, peak value, slope, skewness, and kurtosis, are typically computed from fNIRS signals to train machine learning classifiers [2], [4], [5]. Support vector machine (SVM), linear discriminant analysis (LDA), and artificial neural network (ANN) are classical algorithms. These classifiers mainly rely on hand-crafted features, whereas deep learning can achieve superior classification performance due to powerful feature representation capabilities. Although deep models have made progress in fNIRS classification [6]–[9], numerous barriers hinder in-depth research. First, a large-scale fNIRS signal acquisition may be challenging due to the high cost of fNIRS equipment. Another important reason is that subjects must endure burdensome signal acquisition processes. Additionally, some datasets are protected by privacy policies, making it difficult to obtain permissions for public use. Numerous deep networks are trained on a limited dataset that contains only about ten subjects. Therefore, model generalization performance is not confident enough. Second, differences in evaluation criteria and experimental settings make fair comparisons impossible. For instance, Bak *et al.* [10] used leave-one-out cross-validation (LOO-CV) to train SVM for a single subject. Nazeer *et al.* [11] utilized leave-one-trial-out cross-validation (LOTO-CV) to evaluate the classification performance. Although these evaluation methods have been successfully applied to smaller test sets, they may not be ideal strategies for deep learning models. Third, it is challenging

to directly apply the latest deep learning techniques from computer vision (CV) and natural language processing (NLP) to fNIRS because fNIRS data is different from these domains.

It is reasonable for the first barrier to conduct experiments on larger open-access datasets. We also employ more appropriate evaluation strategies for the second barrier. K-fold cross-validation (KFold-CV) is used to evaluate test results, whereas leave-one-subject-out cross-validation (LOSO-CV) is used to assess model generalization and individual differences [9], [12], [13]. For LOSO-CV, one subject's data serves as a test set, and the rest serves as a training set. The process is repeated until all subjects have been tested. We propose a vision fNIRS framework to address the third barrier. The framework utilizes the Gramian angular difference field (GADF) [14] to convert multi-channel fNIRS signals into multi-channel virtual images. If the signal classification is transformed into an image classification problem, researchers can quickly apply the latest visual models from the CV domain. Therefore, we can tune the hyperparameters of these models in a few minutes, and the classification results are competitive with the latest fNIRS classification models. We directly share CV research with fNIRS classification tasks, which echoes recent NLP-inspired CV developments such as vision Transformers [15]–[17] and masked image modeling [18], [19]. We hope that the findings will inspire more fNIRS studies.

The main contributions are listed as follows:

- We propose a vision fNIRS framework based on GADF to generate multi-channel virtual images for fNIRS classification. Existing visual models can be directly used to train GADF images, and the classification performance is competitive with the latest fNIRS classification models. The proposed framework establishes a bridge between fNIRS and CV.
- We investigate the effects of three different fNIRS signal representations on classification performance, including alternate, stacked, and one-dimensional representations. GADF and Gramian angular summation field (GASF) are two variants of Gramian angular field (GAF). We compare the classification results of sequence-to-image methods, such as GASF, GADF, Markov transition field (MTF), and various combined images. We found that multi-channel GADF images can better encode context dependencies of hemodynamic responses and preserve spatial information.
- Extensive experiments are performed on three open-access datasets. State-of-the-art visual models based on Transformers [15], [16], [20] and multi-layer perceptrons (MLPs) [21], [22] are introduced into comparison experiments. Unlike recent studies [6], [7], [11], [12], our work is more devoted to demonstrating the generality of the proposed framework.

The rest of this paper is organized as follows. Related works are briefly reviewed in Section II. Section III introduces the proposed vision fNIRS framework. Section IV describes open-access datasets and experimental setup. Section V reports the experimental results. Discussion and conclusion are presented in Sections VI and VII, respectively.

II. Related Works

A. fNIRS Classification

Enhancing the classification performance of fNIRS-BCI systems can improve the quality of life for patients who suffer from stroke, spinal cord injury, and amyotrophic lateral sclerosis [6]. Table I summarizes the finger- and foot-tapping tasks. The classification results of these references cannot be compared fairly due to differences in experimental paradigms, subjects, and evaluation criteria. The LOO-CV and LOTO-CV require fewer test samples, which reduces the fidelity and generalization of deep models. By contrast, KFold-CV and LOSO-CV are more suitable for deep learning. In addition, certain experimental results are hardly convincing without extra or larger datasets because deep models are prone to overfitting on limited data.

B. Encoding Time Series to Images

Convolutional neural networks (CNNs) and recurrent neural networks (RNNs) are widely used to process biological signals, such as electrocardiogram (ECG), electroencephalography (EEG), and fNIRS. Multi-channel time-series signals are reshaped into a two-dimensional matrix that resembles an image, then the matrix is fed into a CNN model for classification. Unlike this reshaped signal operation, some sequence-to-image methods, such as GAF [14], MTF [14], and recurrence plot (RP) [23], can encode one-dimensional signals into two-dimensional images. GAF and RP encode EEG signals into image representations for drowsiness detection [24]. Xiao *et al.* [25] extracted features from ECG images generated by GAF to classify hand movements. A CNN is developed for ternary fNIRS classification using GASF [26]. Since few studies apply GAF to fNIRS, we conduct more in-depth studies.

C. Visual Models Based on Transformers and MLPs

Transformer [27], a novel network structure based on the self-attention mechanism, has achieved success in NLP. The significant achievements of Transformers have sparked tremendous interest in CV. Vision Transformer (ViT) [15] is the first image classification model based on pure Transformers and achieves superior performance on large-scale datasets. It splits an image $x \in \mathbb{R}^{H\times W\times C}$ into a sequence of flattened patches $x_p \in \mathbb{R}^{N\times(P^2\times C)}$, where (H, W) is the image resolution, C is the color channel, (P, P) is the patch size, and $N = \frac{HW}{P^2}$. After that, x_p is projected to the model dimension D by a linear layer, named patch embeddings. A positional encoding is added to the embeddings to retain positional information. Then, a learnable classification token $[CLS]$ is appended at the beginning of the embeddings. Finally, the embeddings are fed to Transformer encoders for classification. ViT facilitates follow-up studies [16], [17], [20].

Tolstikhin *et al.* [21] proposed a simple and efficient MLP-Mixer consisting of channel-mixing MLPs and token-mixing MLPs. The channel-mixing MLPs facilitate communication between different channels, and the token-mixing MLPs

TABLE I
SUMMARY OF THE FINGER- AND FOOT-TAPPING TASKS

Reference	Task	Subject	Channel	Method	Evaluation	Accuracy (%)
Bak *et al.* [10]	Right-hand finger-tapping Left-hand finger-tapping Foot-tapping	30	20	SVM	LOO-CV	70.4 ± 18.4
Wang *et al.* [9]	Right-hand finger-tapping Left-hand finger-tapping Foot-tapping	30	20	Transformer	KFold-CV LOSO-CV	75.49 ± 2.07 (KFold-CV) 78.22 ± 16.69 (LOSO-CV)
Nazeer *et al.* [11]	Right-hand finger tapping Left-hand finger tapping Rest	7	24	LDA with vector-based phase analysis	LOO-CV LOTO-CV	85.4 ± 1.4 (LOO-CV) 88.7 ± 2.7 (LOTO-CV)
Sommer *et al.* [12]	Finger tapping with three different frequencies (rest, 80bpm, and 120bpm)	12	12	CNN-LSTM with multi-labeling	LOSO-CV	81
Trakoolwilaiwan *et al.* [6]	Rest, right-, and left-hand motor execution	8	34	CNN	KFold-CV	92.68

allow communication between other spatial locations. MLP-Mixer has aroused interest in modern MLP models [22], [28].

III. METHODS

A. Multi-Channel Virtual Image Generation

The Gramian angular field (GAF) [14] is a novel method to encode time-series signals into image representations, including Gramian angular summation field (GASF) and Gramian angular difference field (GADF). fNIRS signals contain many channels, and the number of channels is dependent on acquisition equipment and experimental paradigms. Specifically, given an fNIRS matrix $X \in \mathbb{R}^{2C\times S}$, where C is the number of channels, S is the number of sampling points, and "2" means two chromophores (HbO and HbR). The sampling point is $S = f \times T$, where f is the sampling frequency and T is the sampling time. Each channel-level HbO or HbR $X_i = \{x_1, x_2, \ldots, x_S\}$ is scaled to the interval $[-1, 1]$:

$$\widetilde{x_j} = \frac{\left(x_j - \max\left(X_i\right)\right) + \left(x_j - \min\left(X_i\right)\right)}{\max\left(X_i\right) - \min\left(X_i\right)}, \tag{1}$$

where $\widetilde{x_j}$ is the normalized sampling point, $i = 1, 2, \ldots, C$ and $j = 1, 2, \ldots, S$. The rescaled channel-level signal $\widetilde{X_i}$ is transformed from the Cartesian coordinates to the polar coordinate system by

$$\begin{cases} \phi_j = \cos^{-1}\left(\widetilde{x}_j\right), -1 \le \widetilde{x_j} \le 1, \widetilde{x_j} \in \widetilde{X_i} \\ r_j = \dfrac{t_j}{N}, t_j \in \mathbb{N} \end{cases} \tag{2}$$

where t_j is the time stamp and N is a constant factor to regularize the span of the polar coordinate system. Then, GASF and GADF are defined as follows:

$$GASF_i = \begin{bmatrix} \cos\left(\phi_1 + \phi_1\right) & \cdots & \cos\left(\phi_1 + \phi_S\right) \\ \vdots & \ddots & \vdots \\ \cos\left(\phi_S + \phi_1\right) & \cdots & \cos\left(\phi_S + \phi_S\right) \end{bmatrix}, \tag{3}$$

$$GADF_i = \begin{bmatrix} \sin\left(\phi_1 - \phi_1\right) & \cdots & \sin\left(\phi_1 - \phi_S\right) \\ \vdots & \ddots & \vdots \\ \sin\left(\phi_S - \phi_1\right) & \cdots & \sin\left(\phi_S - \phi_S\right) \end{bmatrix}. \tag{4}$$

Finally, the $GASF_i$ or $GADF_i$ matrix is stacked along the depth dimension to generate a multi-channel virtual image $\widetilde{X} \in \mathbb{R}^{S\times S\times 2C}$.

In general, S is a larger value that requires huge computational resources for model training. Piecewise aggregation approximation (PAA) is used to compress fNIRS signals while keeping time series trends [29]. For the previously mentioned channel-level signal $X_i = \{x_1, x_2, \ldots, x_S\}$, let M be the dimension of the fNIRS sequence $\overline{X_i} = \{\overline{x_1}, \overline{x_2}, \ldots, \overline{x_M}\}$ worked with $(1 \le M \le S)$. The m^{th} element $\overline{x_m}$ of $\overline{X_i}$ is calculated by

$$\overline{x_m} = \frac{M}{S} \sum_{j=\frac{S}{M}(m-1)+1}^{\frac{S}{M}m} x_j. \tag{5}$$

In the vision fNIRS framework, the PAA compression size M is set to 32 to achieve a trade-off between performance and speed. Then, $\overline{X_i}$ is encoded into a multi-channel virtual image.

B. Vision fNIRS Framework

We propose a vision fNIRS framework to overcome domain barriers between fNIRS and CV. As illustrated in Fig. 1, the proposed framework consists of three modules: data preprocessing, image generation, and visual model training.

1) Data Preprocessing: Extracting features directly from raw signals may impair classification performance due to the inherent noise in raw fNIRS signals. Typically, data preprocessing consists of the modified Beer-Lambert law (MBLL) [30], filtering, segmentation, and baseline correction. MBLL calculates the concentration changes of HbO and HbR because the chromophores have different absorption coefficients at different near-infrared wavelengths. High-pass and band-pass filters are applied to fNIRS signals to reduce noise and artifacts, and baseline correction is used to correct for baseline drift. Finally, fNIRS signals are divided into sequences of predetermined length for classification. In general, the preprocessing for different datasets is inconsistent. As a result, the specific preprocessing steps should be consistent with the original literature.

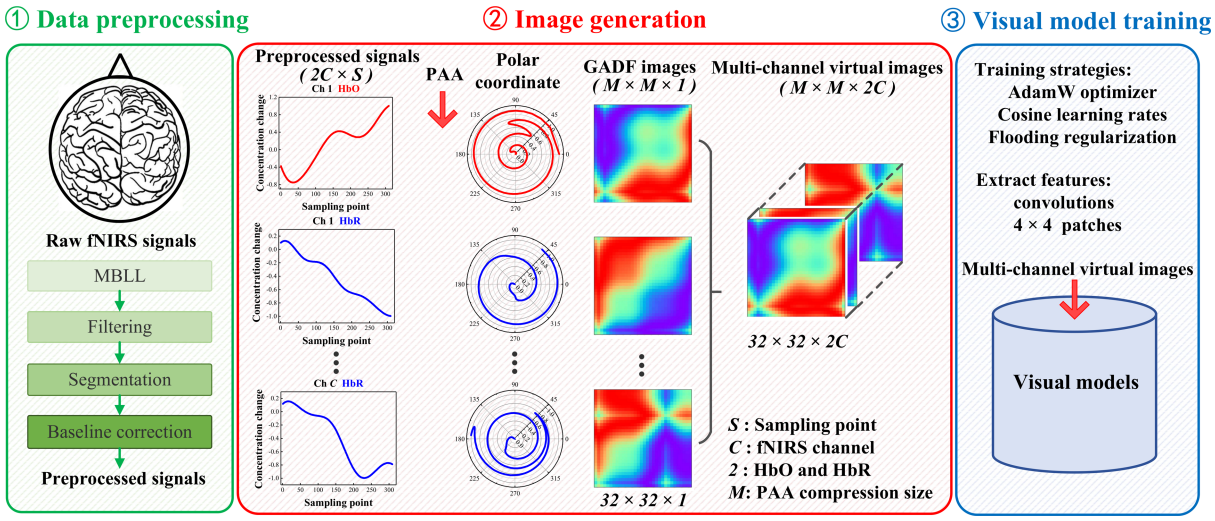

Fig. 1. Overview of vision fNIRS framework. The framework consists of three modules: data preprocessing, image generation, and visual model training. Data preprocessing is a crucial step for classification due to the inherent noise in raw fNIRS signals. Image generation serves as a bridge for visual models to train fNIRS signals directly. Finally, advanced training strategies and visual models are used to classify multi-channel virtual images encoded by GADF.

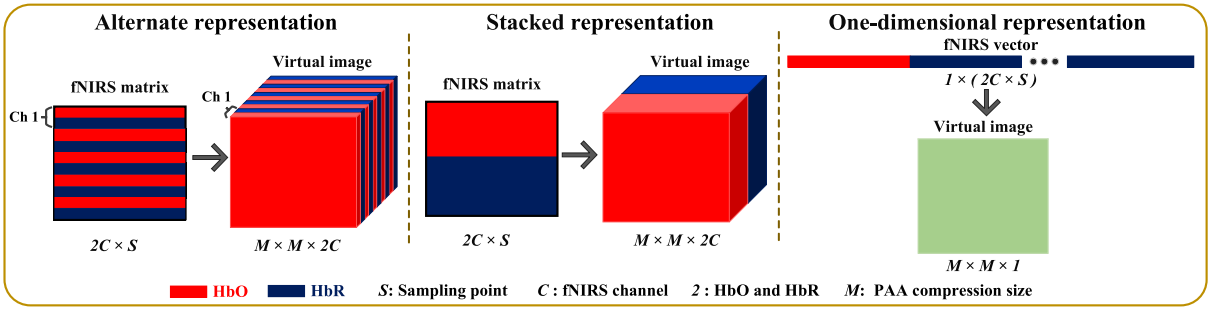

Fig. 2. Three different fNIRS signal representations. The vision fNIRS framework adopts the alternate representation by default.

2) Image Generation: Image generation is the foundation of the vision fNIRS framework. The size of fNIRS signals may be variable due to acquisition equipment and experimental paradigms. PAA can compress fNIRS signals to a predetermined length and keep signal trends. In the framework, each channel signal is compressed to 32 by PAA, and GADF encodes the temporal correlation of hemodynamic responses to a $32 \times 32 \times 1$ image. Each GADF image is stacked along the depth dimension to form a multi-channel virtual image to preserve spatial information. Finally, visual models extract GADF image features through convolutions or 4×4 patches. Note that the framework only uses GADF instead of GASF because GADF is more effective in our experiments. The size $(32 \times 32 \times 1)$ of GADF images is consistent with CIFAR [31] (except for color channels). Therefore, we can transplant a series of visual networks from CV to fNIRS, including the popular ResNet and the latest ViT- and MLP-like architectures.

3) Visual Model Training: A simple training strategy that incorporates AdamW optimizer [32], cosine learning rates [33], and flooding regularization [38] is developed. Overfitting degrades classification performance and model generalization due to the limited amount of fNIRS data. Although many fNIRS classification models address this problem by fine-tuning dropout rates and weight decay, the test loss keeps increasing as the training loss tends to zero. Owing to these methods cannot solve overfitting directly, a simple and efficient flooding regularization is used for our framework. Flooding directly limits the training loss around a small constant value called flooding level rather than a zero loss [34]. Given the cross-entropy loss function $L(y_i, \widehat{y}_i)$:

$$L(y_i, \widehat{y}_i) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log \widehat{y}_i + (1 - y_i) \log (1 - \widehat{y}_i), \tag{6}$$

the modified $L(y_i, \widehat{y}_i)$ with flooding is

$$\tilde{L}(y_i, \widehat{y}_i) = |L(y_i, \widehat{y}_i) - b| + b, \tag{7}$$

where y_i is the true label, $\widehat{y}_i$ is the predicted output, and b is the flooding level.

C. fNIRS Signal Representations

We are interested in whether different fNIRS signal representations affect the classification performance of visual models under the vision fNIRS framework. To the best of our knowledge, this is the first study on the question. Fig. 2 illustrates three typical representations. The alternate representation is described as an alternating combination of HbO and

HbR for each channel. The stacked representation combines the same chromophores (HbO and HbR). The one-dimensional representation reshapes fNIRS signals into a one-dimensional vector before generating a single-channel GADF image. The alternate and stacked representations can reflect the effects of the spatial location of HbO and HbR on classification performance. The one-dimensional representation can determine whether a single-channel image has an advantage over a multi-channel image. In Section V, we compare the performance of these representations. The vision fNIRS framework adopts the alternate representation by default unless otherwise specified.

IV. Experiments

A. Open-Access Datasets

1) MA: In the mental arithmetic (MA) experiment, 29 subjects (14 males and 15 females, average age 28.5 ± 3.7 years) participated in the study [35]. For MA tasks, the subjects were asked to remember an initial subtraction that appeared on the screen. They repeatedly performed to subtract the one-digit number from the previous subtraction during the 10 s task period. For the baseline (BL) task, they were instructed to take a rest. They were asked to rest during the 15–17 s rest period. The dataset is available at http://doc.ml.tu-berlin.de/hBCI.

2) WG: A total of 26 volunteers (9 males and 17 females, average age 26.1 ± 3.5 years) participated in word generation (WG) tasks [36]. During the 10 s period of WG tasks, the participants were instructed to continue thinking about the first letter of the word shown on the preceding screen as soon as possible. For the BL task, they were required to relax and gaze at the fixation cross to maintain a low cognitive load. The participants were asked to relax to avoid excessive eye movements during the rest period. The dataset is available at http://doc.ml.tu-berlin.de/simultaneous_EEG_NIRS/.

3) UFFT: A total of 30 volunteers (17 males and 13 females, 23.4 ± 2.5 years old) participated in motor execution tasks, including right-hand finger-tapping (RHT), left-hand finger-tapping (LHT), and foot-tapping (FT) [10]. During the task, they conducted a specific movement according to the instruction randomly displayed on the screen. The dataset is available at https://doi.org/10.6084/m9.figshare.9783755.v1.

MA and WG are two-class classification tasks (MA vs. BL and WG vs. BL, respectively), and UFFT is a three-class classification task (RHT vs. LHT vs. FT). The data preprocessing is consistent with the original literature. Finally, considering delayed hemodynamic responses [37], [38], the segmented sample point windows for MA, WG, and UFFT are 80–320, 80–300, and 20–320, respectively.

B. Experimental Setup

The CNN [39], CNN-3b [40], RNN [39], fNIRS-T [9], and fNIRS-PreT [9] are used as baseline networks. A three-branch CNN network decodes consumers' preference levels from viewing commercial advertisement videos of different durations (15, 30, and 60 s) [40]. We use the subnetwork (CNN-3b) developed for a 15 s video branch because the 15 s video is close to the task time of the open-access datasets. ResNet-18 [41], ViT [15], EarlyConvViT [20], PVTv2-B0 [16], MLP-Mixer [21], and ResMLP [22] are chosen for comparison experiments. These visual models are trained by AdamW (default parameters) with a batch size of 128 for 150 epochs. The maximum number of iterations T_{max} for cosine learning rates is set to 30. The configuration of the visual models is listed in Table II. The piecewise decay flooding [9] requires more empirical tuning, so a fixed flooding level is used for simplicity. The flooding level is reported in Table III. We conduct 5 × 5-fold cross-validation (KFold-CV) experiments to evaluate the accuracy (mean ± std). In subject-independent experiments, LOSO-CV is used to evaluate model generalization and individual differences. The accuracy, precision, recall, F1-score, and kappa coefficient are used as overall performance metrics (macro-average for UFFT). The vision fNIRS framework is implemented by PyTorch [42] and trained on NVIDIA GeForce GTX 1080 and Tesla V100 GPUs.

TABLE II
Configuration of Visual Models. The MLP Size of MLP-Mixer and ResMLP Is Hidden Size × Expansion Factor (4 by Default)

Model	Patch size	Layers	Hidden size	MLP size	Heads
ResNet-18	-	-	-	-	-
ViT	4 × 4	8	128	64	8
EarlyConvViT	-	6	128	64	8
PVTv2-B0	4 × 4	8	-	-	-
MLP-Mixer	4 × 4	8	128	128 × 4	8
ResMLP	4 × 4	8	128	128 × 4	8

TABLE III
Flooding Level for Different Models

Domain	Model	MA	WG	UFFT
fNIRS	CNN	0.25	0.25	0.30
	CNN-3b	0.40	0.30	0.00
	RNN	0.60	0.65	0.60
	fNIRS-T	-	0.35	-
	fNIRS-PreT	-	0.35	-
CV	ResNet-18	0.25	0.25	0.25
	ViT	0.20	0.20	0.20
	EarlyConvViT	0.20	0.20	0.20
	PVTv2-B0	0.20	0.20	0.20
	MLP-Mixer	0.20	0.20	0.20
	ResMLP	0.20	0.25	0.20

TABLE IV
Average Accuracy (%) of KFold-CV for the Alternate Representation. Bold Values Highlight the Best Results for Each Dataset

Domain	Model	KFold-CV		
		MA	WG	UFFT
fNIRS	CNN	75.89 ± 2.19	70.87 ± 1.93	74.49 ± 1.20
	CNN-3b	72.79 ± 2.90	69.53 ± 2.33	73.70 ± 1.69
	RNN	61.29 ± 1.97	59.95 ± 1.77	61.10 ± 1.59
	fNIRS-T	76.59 ± 1.86	72.36 ± 2.04	**75.49 ± 2.07**
	fNIRS-PreT	78.53 ± 1.83	72.14 ± 1.48	70.66 ± 1.45
CV	ResNet-18	78.57 ± 1.55	73.29 ± 1.73	72.94 ± 1.64
	ViT	77.49 ± 1.86	71.42 ± 2.13	73.17 ± 1.66
	EarlyConvViT	77.47 ± 1.53	72.67 ± 1.85	72.87 ± 1.26
	PVTv2-B0	**78.68 ± 1.59**	73.00 ± 1.41	72.68 ± 1.43
	MLP-Mixer	77.98 ± 1.69	**73.92 ± 1.45**	74.14 ± 1.39
	ResMLP	77.74 ± 1.65	73.23 ± 1.62	74.36 ± 1.90

TABLE V
RESULTS OF LOSO-CV FOR THE ALTERNATE REPRESENTATION. BOLD VALUES HIGHLIGHT THE BEST RESULTS FOR EACH EVALUATION METRIC

Dataset	Domain	Method	LOSO-CV				
			Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)	Kappa
MA	fNIRS	CNN	76.21 ± 7.34	78.27 ± 9.76	75.40 ± 13.52	75.60	0.52
		CNN-3b	74.89 ± 7.11	79.63 ± 10.14	69.66 ± 12.01	73.28	0.50
		RNN	67.47 ± 6.63	68.30 ± 8.09	67.59 ± 12.83	67.11	0.35
		fNIRS-T	78.16 ± 7.59	79.49 ± 8.35	76.90 ± 11.97	77.59	0.56
		fNIRS-PreT	80.57 ± 7.53	80.91 ± 9.66	**82.30 ± 12.13**	80.67	0.61
	CV	ResNet-18	**81.44 ± 8.16**	**82.97 ± 9.47**	80.00 ± 10.69	**81.05**	**0.63**
		ViT	79.60 ± 9.66	80.41 ± 10.77	79.54 ± 11.30	79.55	0.59
		EarlyConvViT	79.60 ± 8.82	80.44 ± 11.09	80.23 ± 11.58	79.60	0.59
		PVTv2-B0	81.09 ± 8.98	81.56 ± 10.41	81.72 ± 12.31	81.00	0.62
		MLP-Mixer	80.63 ± 8.34	81.58 ± 9.66	80.92 ± 13.42	80.41	0.61
		ResMLP	79.77 ± 8.77	79.94 ± 9.92	81.72 ± 12.58	80.03	0.60
WG	fNIRS	CNN	74.36 ± 6.98	75.99 ± 10.66	74.10 ± 7.75	74.42	0.49
		CNN-3b	72.05 ± 8.04	71.42 ± 11.57	**80.77 ± 13.63**	74.25	0.44
		RNN	65.45 ± 6.69	66.97 ± 7.52	63.33 ± 15.58	63.83	0.31
		fNIRS-T	76.15 ± 7.81	77.29 ± 11.40	76.79 ± 7.31	76.52	0.52
		fNIRS-PreT	76.35 ± 7.38	75.87 ± 8.28	78.21 ± 10.55	76.64	0.53
	CV	ResNet-18	76.09 ± 6.44	75.88 ± 8.27	78.97 ± 10.08	76.70	0.52
		ViT	74.23 ± 5.55	73.34 ± 8.27	79.10 ± 11.82	75.21	0.48
		EarlyConvViT	75.51 ± 5.79	75.05 ± 8.61	78.85 ± 7.39	76.39	0.51
		PVTv2-B0	**76.99 ± 5.70**	77.67 ± 8.16	77.56 ± 7.49	77.15	**0.54**
		MLP-Mixer	76.92 ± 6.27	**77.87 ± 9.70**	77.95 ± 9.25	**77.16**	**0.54**
		ResMLP	75.58 ± 7.20	75.34 ± 9.10	78.59 ± 9.44	76.33	0.51
UFFT	fNIRS	CNN	76.67 ± 16.44	78.21 ± 15.65	76.67 ± 16.44	76.21	0.65
		CNN-3b	75.64 ± 16.76	77.64 ± 15.91	75.64 ± 16.76	74.93	0.63
		RNN	65.51 ± 15.57	66.11 ± 15.92	65.51 ± 15.57	64.48	0.48
		fNIRS-T	**78.22 ± 16.69**	**79.36 ± 16.67**	**78.22 ± 16.69**	**77.59**	**0.67**
		fNIRS-PreT	76.98 ± 16.35	77.66 ± 16.09	76.98 ± 16.35	76.72	0.65
	CV	ResNet-18	76.53 ± 15.70	77.20 ± 15.54	76.53 ± 15.70	75.95	0.65
		ViT	77.82 ± 15.57	78.28 ± 15.47	77.82 ± 15.57	77.37	**0.67**
		EarlyConvViT	76.58 ± 15.66	77.53 ± 15.12	76.58 ± 15.66	76.30	0.65
		PVTv2-B0	76.27 ± 15.17	77.30 ± 14.57	76.27 ± 15.17	75.82	0.64
		MLP-Mixer	76.98 ± 15.72	77.88 ± 15.17	76.98 ± 15.72	76.55	0.65
		ResMLP	76.98 ± 16.67	78.14 ± 16.09	76.98 ± 16.67	76.35	0.65

V. RESULTS

A. Classification Results

We compared many models from fNIRS and CV. Table IV presents the average accuracy of KFold-CV for the alternate representation. Among the baseline models, Transformer-based fNIRS-T and fNIRS-PreT outperform CNN, CNN-3b, and RNN. It is widely accepted that RNNs are suitable for sequence modeling, while our experiments demonstrate that the classification performance is worse than other networks (Wilcoxon signed-rank test, $p < 0.001$). The RNN achieved just about 60% classification accuracy on three datasets. The possible reason is that hemodynamic responses lack distinct sequence patterns in the inactive area, preventing RNN from capturing vital contextual information. For MA tasks, ResNet-18 and PVTv2-B0 are competitive with fNIRS-PreT and significantly outperform the other baseline models (Wilcoxon signed-rank test, $p < 0.001$). Although visual models perform marginally worse than fNIRS-T on UFFT, the average accuracy of the five visual models exceeds fNIRS-T on WG. MLP-Mixer achieved the highest classification results of 73.92%. Experimental results validate the effectiveness of the proposed framework.

We explain why GADF images improve classification performance in terms of temporal-spatial features, amount of information, and inductive biases. GADF encodes both local temporal correlation and long-term context dependencies into an image, and multi-channel spatial information is transformed to the channel dimension of virtual images. Visual models can easily extract temporal and spatial features from the multi-channel GADF images. The input data types for classification mainly include hand-crafted features, preprocessed signals, and virtual images. For example, if there are 20 channels and each channel contains 200 sampling points, then the number of input data is $20 \times 2 \times 6 = 240$ (2 chromophores and 6 statistical features), $20 \times 2 \times 200 = 8000$, and $20 \times 2 \times 32 \times 32 = 40960$, respectively. Compared with signal-based inputs, virtual images increase the amount of information by a factor of 5. The GADF images can increase information density from original signals, which helps to train deep models adequately. It is commonly understood that Transformer-based models lack convolutional inductive biases and require more data to perform better [15], [20]. However, we do not observe this phenomenon in the experiments. This suggests the GADF images compensate for inductive biases to some extent rather than directly extracting features from fNIRS signals.

B. Subject-Independent Experiments

Individual differences exist among subjects because of life background, task collaboration, and response sensitivity. Subject-independent experiments can verify model generalization that evaluates the performance of a model on other scenarios or subjects. Table V reports the LOSO-CV results for

TABLE VI
AVERAGE ACCURACY (%) OF KFOLD-CV FOR THE STACKED REPRESENTATION. THE NUMBER IN PARENTHESIS MEANS THE *p*-VALUE (WILCOXON SIGNED-RANK TEST) BETWEEN THE ALTERNATE AND STACKED REPRESENTATIONS

Model	KFold-CV		
	MA	WG	UFFT
ResNet-18	78.30±1.31 (0.432)	72.47±1.70 (0.126)	72.46±1.49 (0.203)
ViT	77.47±1.47 (0.855)	70.77±2.36 (0.331)	73.48±1.33 (0.533)
EarlyConvViT	77.40±1.93 (0.871)	72.45±2.26 (0.587)	72.90±1.78 (0.979)
PVTv2-B0	78.34±1.46 (0.331)	73.27±1.57 (0.560)	72.69±1.84 (0.867)
MLP-Mixer	78.20±1.71 (0.704)	73.99±1.29 (0.979)	74.08±1.21 (0.874)
ResMLP	78.02±2.05 (0.988)	73.40±1.74 (0.879)	74.51±1.75 (0.943)

TABLE VII
AVERAGE ACCURACY (%) OF KFOLD-CV FOR THE ONE-DIMENSIONAL REPRESENTATION ON MA

Model	Size (H, W, C)	Params (M)	FLOPs (M)	Accuracy (%)
EarlyConvViT	(32, 32, 72)†	2.7	24.7	77.47 ± 1.53
	(32, 32, 72)‡	2.7	24.7	77.47 ± 1.47
	(32, 32, 1)	2.6	16.8	68.58 ± 1.71
	(64, 64, 1)	2.6	63.4	73.45 ± 2.22
	(128, 128, 1)	2.6	267.4	73.22 ± 2.06
	(256, 256, 1)	2.6	1366.5	75.52 ± 1.66
PVTv2-B0	(32, 32, 72)†	3.5	17.8	78.68 ± 1.59
	(32, 32, 72)‡	3.5	17.8	78.34 ± 1.46
	(32, 32, 1)	3.4	10.7	68.97 ± 2.55
	(64, 64, 1)	3.4	43.0	73.05 ± 1.74
	(128, 128, 1)	3.4	175.1	73.33 ± 1.41
	(256, 256, 1)	3.4	749.2	73.05 ± 1.73

[1] Note: The superscript † and ‡ indicate the results of the alternate and stacked representations, respectively.

the alternate representation. The visual models achieve better average classification accuracy and generalization ability on MA. The accuracy and F1-score of ResNet-18 both exceeded 81%. PVTv2-B0 and MLP-Mixer outperform fNIRS-T and fNIRS-PreT in all evaluation metrics on the WG task. Visual models achieve the highest F1-score on MA and WG. For UFFT, visual models obtain competitive results with fNIRS-T and fNIRS-PreT. Although fNIRS-T achieves higher average accuracy, the F1-score obtained by most visual models is competitive with fNIRS-T. It means that the performance gap may not be significant. In addition, the higher kappa coefficients demonstrate the superiority of the vision fNIRS framework.

C. Results of Signal Representations

In this section, we investigate the effects of signal representations on classification performance. Table VI shows the average accuracy of KFold-CV for the stacked representation. The classification results of the alternate and stacked representations are not significantly different. The representations can increase information density and allow convolutional layers and self-attention to extract temporal and spatial features efficiently. However, the one-dimensional representation is worse than the previous representations. PAA substantially compresses multi-channel long-range contextual information into a one-dimensional vector, such as activation patterns and hemodynamic responses. More importantly, it is hard for one-dimensional representation to preserve the spatial information of fNIRS signals from a one-dimensional vector. Table VII summarizes the quantitative results of KFold-CV for EarlyConvViT and PVTv2-B0 on MA. While increasing the image size from 32 to 256 may reduce information missing, it dramatically increases computational costs and gradually saturates the classification results. When the image size of PVTv2-B0 is set to 128, the one-dimensional representation is 5.35% and 5.01% lower than the alternate (78.68%) and stacked (78.34%) representations, respectively. EarlyConvViT with an input size of 256 is 1.95% lower than the alternate and stacked representations and increases FLOPs by 55 times. Therefore, the alternate and stacked representations are more practical.

D. Comparison With Other Virtual Images

We further compare other virtual images, including GASF, MTF, GASF-MTF, GADF-MTF, and GADF-GASF. The results on UFFT are shown in Table VIII. The vision fNIRS framework uses GADF instead of GASF. The main difference between (3) and (4) is the encoding function. Surprisingly, this little distinction leads to a considerable performance discrepancy. Compared with GADF images (see Table IV), the average accuracy of these visual models has dropped significantly. Some studies also found that GADF can obtain more accurate classification results than GASF [25], [43]. Xiao *et al.* [25] found that GASF images perform somewhat worse in detail compared to GADF images. We attempt to explain reason from the perspective of fNIRS signal characteristics. In the Euclidean space, the inner product measures the similarity of two vectors u and v, and it is defined as

$$\langle u, v \rangle = \|u\| \cdot \|v\| \cdot \cos(\theta). \tag{8}$$

In (3), GASF encodes the cosine similarity of each pair of time intervals [44]. The main characteristic of fNIRS signals is the delayed hemodynamic response [37], [38] that causes no significant changes in several adjacent sampling points of HbO and HbR. GASF may not efficiently encode the local information of hemodynamic responses. In contrast, GADF considers the trigonometric difference of a pair of time intervals that can better capture the change in adjacent hemoglobin concentration. It is essential for convolutions and self-attention to extract local features and global dependencies. For MA tasks, the GADF and GASF images for Subject 1 are shown in Fig. 3. Many low-brightness areas appear in the GASF images compared to the GADF images, which indicates that the feature information richness of the GASF images may be lower than the GADF images.

MTF encodes time-series information by representing the first-order Markov transition probability [14]. Compared with GADF (see Table IV), MTF performs poorly in classification tasks. In addition, MTF is more prone to overfitting than GADF due to the uncertainty in MTF inverted image [14]. Therefore, it seems that the MTF images are unsuitable for classification tasks.

We assess the effects of combined images on classification, including GASF-MTF, GADF-MTF, and GADF-GASF.

TABLE VIII
AVERAGE ACCURACY (%) OF KFOLD-CV ON UFFT. THE NUMBER IN PARENTHESIS INDICATES DECREASED ACCURACY OVER THE GADF IMAGES

Model	KFold-CV				
	GASF	MTF	GASF-MTF	GADF-MTF	GADF-GASF
ResNet-18	54.23 ± 1.61 (-18.71)	51.82 ± 1.04 (-21.12)	54.31 ± 1.34 (-18.63)	72.58 ± 1.20 (-0.36)	70.98 ± 1.77 (-1.96)
ViT	56.67 ± 1.76 (-16.50)	53.78 ± 2.19 (-19.39)	54.00 ± 1.71 (-19.17)	71.87 ± 2.07 (-1.30)	71.78 ± 1.27 (-1.39)
EarlyConvViT	53.65 ± 1.64 (-19.22)	46.98 ± 1.60 (-25.89)	53.16 ± 1.19 (-19.71)	72.93 ± 2.37 (+0.06)	71.64 ± 0.86 (-1.23)
PVTv2-B0	51.94 ± 1.62 (-20.74)	45.56 ± 1.16 (-27.12)	52.58 ± 1.10 (-20.10)	71.51 ± 1.55 (-1.17)	70.98 ± 1.61 (-1.70)
MLP-Mixer	56.43 ± 1.58 (-17.71)	49.11 ± 1.38 (-25.03)	55.87 ± 2.72 (-18.27)	73.60 ± 1.90 (-0.54)	72.49 ± 2.05 (-1.65)
ResMLP	58.44 ± 2.00 (-15.92)	54.76 ± 1.30 (-19.60)	58.71 ± 2.27 (-15.65)	73.38 ± 2.39 (-0.98)	73.73 ± 0.81 (-0.63)

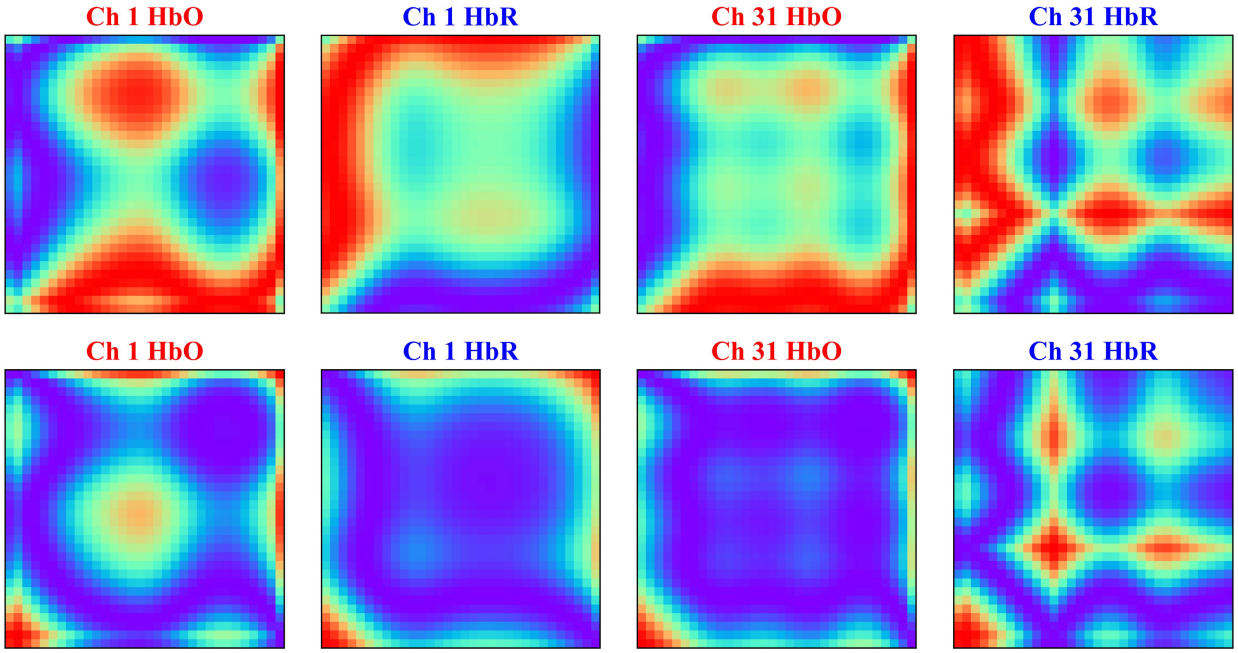

Fig. 3. For the MA task of Subject 1, the fNIRS signals at Channels (Ch) 1 and 31 are encoded as virtual images by GADF (top row) and GASF (bottom row). Many low-brightness areas appear in the GASF images compared to the GADF images.

TABLE IX
DIFFERENT PATCH SIZES FOR VIT AND MLP-MIXER ON MA. THE PATCH SIZE MUST BE DIVISIBLE BY A 32 × 32 GADF IMAGE

Model	Patch size ($P \times P$)	Params (M)	FLOPs (M)	Accuracy (%)
ViT	2 × 2	2.3	1125.8	79.29 ± 1.69
	4 × 4	2.4	189.5	77.49 ± 1.86
	8 × 8	2.8	49.9	74.41 ± 2.10
	16 × 16	4.6	20.8	71.62 ± 2.33
MLP-Mixer	2 × 2	5.3	817.5	78.10 ± 2.13
	4 × 4	1.5	110.8	77.98 ± 1.69
	8 × 8	1.7	28.5	76.59 ± 2.47
	16 × 16	3.4	13.8	73.07 ± 2.42

TABLE X
ABLATION STUDY ON THE UFFT DATASET

Input	Model	Params (M)	FLOPs (M)	Accuracy (%)
GADF	PVTv2-B0	3.5	14.6	72.68 ± 1.43
	MLP-Mixer	1.4	106.6	74.14 ± 1.39
	ResMLP	1.2	76.5	74.36 ± 1.90
Signal	PVTv2-B0	3.4	148.7	71.78 ± 0.73
	MLP-Mixer	37.1	5399.8	70.67 ± 1.30
	ResMLP	5.6	1364.0	73.87 ± 1.07

An obvious disadvantage for these combined images is the ever-increasing computational costs as the number of image channels increases from 40 to 80. Since MTF and GASF have a lower classification performance than GADF, MTF and GASF may interfere with models to extract meaningful features from GADF images. It may be the reason for the performance degradation of combined images.

E. Patch Size

The patch size is a crucial hyperparameter that affects model performance and training costs. ViT and MLP-Mixer are trained with different patch sizes to illustrate these problems. The results of KFold-CV on MA are shown in Table IX. As the patch size of ViT is reduced from 16 × 16 to 2 × 2, the classification accuracy keeps improving, but huge FLOPs require more computational resources and training time. The patch operation for input images fails to preserve the important local information among neighboring patches. A smaller patch can alleviate local information loss but significantly increases computational costs. We also observe a similar situation on MLP-Mixer. Therefore, the 4 × 4 patch size is more in line with the trade-off between performance and speed.

VI. DISCUSSION

The primary motivation of our study is to improve fNIRS classification tasks by leveraging rich achievements from

the CV domain. We propose a GADF-based vision fNIRS framework to transform multi-channel fNIRS signals into multi-channel virtual images. GADF encodes the short-term and long-term dependencies of fNIRS signals, corresponding to the changes in hemoglobin concentration and hemodynamic responses, respectively. Multi-channel GADF images keep the temporal correlation and spatial relationships. Considering two distinct types of chromophores (HbO and HbR), alternate, stacked, and one-dimensional representations are discussed. Visual models from the CV domain can be directly used for fNIRS classification. Many studies [6]–[8], [12] use simple reshaping operations to stack multi-channel fNIRS signals into a two-dimensional matrix, but the essence of the matrix is still time-series data. The main purpose of this operation is to match the input form of CNNs rather than the properties of fNIRS signals. Therefore, multi-channel virtual image generation is a more advanced and effective method.

Extensive experiments demonstrate the superiority of the framework. In Table IV, visual models perform slightly worse than fNIRS-T on UFFT, whereas most visual models achieve higher average classification accuracy. PVTv2-B0 and MLP-Mixer achieved the highest classification accuracy of 78.68% and 73.92% on MA and WG, respectively. Compared with other models, the average accuracy of RNN was about 60% on the three datasets. The RNN may be inefficient in capturing features of hemodynamic responses from fNIRS signals. In Table V, more comprehensive evaluation metrics, F1-score and Kappa coefficient, show that the visual models are competitive with fNIRS-T on UFFT. Then, the effectiveness of three fNIRS signal representations is evaluated. In Table VI, the stacked representation is competitive with the alternate representation. However, the one-dimensional representation is worse than the other representations. The reason mainly comes from two aspects. PAA over-compresses the long-range contextual information because multi-channel fNIRS signals are compressed into a one-dimensional vector that is difficult to preserve spatial information. Although increasing the size from 32 to 256 can improve the classification performance, computational costs increase dramatically and the performance exhibits a saturated state. Furthermore, we carefully study the effects of diverse virtual images and their combinations on classification tasks. Table VIII indicates that the GADF images have better classification performance and training efficiency. The poor performance of GASF and MTF also confirms previous studies [14], [25], [43]. Besides, the combination of different virtual images, such as GASF-MTF, GADF-MTF, and GADF-GASF, do not show advantages in our experiments. The patch size is a vital hyperparameter. A smaller patch can alleviate information lost among neighboring patches. The average accuracy of ViT improved from 77.49% to 79.29% when the patch size decreased from 4×4 to 2×2, but FLOPs increased from 189.5M to 1125.8M. Therefore, the 2×2 patch size can achieve better performance under sufficient computing resources; otherwise, the 4×4 patch size is more practical.

Ablation experiments were conducted on the UFFT dataset. Since the number of patches generated by fNIRS signals is much more than GADF images, input fNIRS signals require huge computational resources when they are fed directly to ViT- and MLP-like models. Some experiments exceed the capabilities of our experimental platform, so only partial ablation results are reported in Table X. We can draw two important conclusions:

- The visual models using GADF images can acquire higher average classification accuracy. The ablation results prove that GADF images are more efficient than fNIRS signals.
- Training visual models using fNIRS signals requires huge FLOPs, whereas our framework is more affordable and practical. Existing visual models can be efficiently trained under the framework. Therefore, the proposed framework bridges the gap between fNIRS and CV.

Even though the study is encouraging, our framework still has some limitations. Since these visual models are initially designed for ImageNet [45], we simply modify some hyperparameters without fine-tuning. The true classification performance of these models may be underestimated. Apart from that, we do not design a specialized classification model for the vision fNIRS framework. Our study aims to demonstrate the generality and scalability of the proposed framework rather than a special model architecture.

VII. Conclusion

This paper proposes a general and scalable vision fNIRS framework that uses GADF to encode multi-channel fNIRS signals into multi-channel images. The framework transforms sequence classification into an image classification problem effectively. Existing visual models can be directly used to train multi-channel images and achieve competitive classification performance. Extensive experiments based on three open-access datasets confirm the effectiveness of the proposed framework. Furthermore, the effects of different signal representations on classification performance are discussed. We also analyze the reason that the classification results of GADF are superior to other virtual images. We hope to introduce CV research to improve fNIRS classification studies and inspire future work.

References

[1] F. F. Jöbsis, "Noninvasive, infrared monitoring of cerebral and myocardial oxygen sufficiency and circulatory parameters," *Science*, vol. 198, no. 4323, pp. 1264–1267, 1977.

[2] N. Naseer and K.-S. Hong, "FNIRS-based brain-computer interfaces: A review," *Frontiers Hum. Neurosci.*, vol. 9, p. 3, Jan. 2015.

[3] S. Coyle, T. Ward, C. Markham, and G. McDarby, "On the suitability of near-infrared (NIR) systems for next-generation brain–computer interfaces," *Physiol. Meas.*, vol. 25, no. 4, pp. 815–822, Aug. 2004.

[4] N. Naseer, M. J. Hong, and K.-S. Hong, "Online binary decision decoding using functional near-infrared spectroscopy for the development of brain–computer interface," *Exp. Brain Res.*, vol. 232, no. 2, pp. 555–564, Feb. 2014.

[5] N. Naseer and K.-S. Hong, "Classification of functional near-infrared spectroscopy signals corresponding to the right- and left-wrist motor imagery for development of a brain–computer interface," *Neurosci. Lett.*, vol. 553, pp. 84–89, Oct. 2013.

[6] T. Trakoolwilaiwan, B. Behboodi, J. Lee, K. Kim, and J.-W. Choi, "Convolutional neural network for high-accuracy functional near-infrared spectroscopy in a brain–computer interface: Three-class classification of rest, right-, and left-hand motor execution," *Neurophotonics*, vol. 5, no. 1, p. 011008, 2017.

[7] U. Asgher *et al.*, "Enhanced accuracy for multiclass mental workload detection using long short-term memory for brain–computer interface," *Frontiers Neurosci.*, vol. 14, p. 584, Jun. 2020.

[8] T. Ma *et al.*, "CNN-based classification of fNIRS signals in motor imagery BCI system," *J. Neural Eng.*, vol. 18, no. 5, Oct. 2021, Art. no. 056019.

[9] Z. Wang, J. Zhang, X. Zhang, P. Chen, and B. Wang, "Transformer model for functional near-infrared spectroscopy classification," *IEEE J. Biomed. Health Informat.*, vol. 26, no. 6, pp. 2559–2569, Jun. 2022.

[10] S. Bak, J. Park, J. Shin, and J. Jeong, "Open-access fNIRS dataset for classification of unilateral finger- and foot-tapping," *Electronics*, vol. 8, no. 12, p. 1486, Dec. 2019.

[11] H. Nazeer *et al.*, "Enhancing classification accuracy of fNIRS-BCI using features acquired from vector-based phase analysis," *J. Neural Eng.*, vol. 17, no. 5, Oct. 2020, Art. no. 056025.

[12] N. M. Sommer, B. Kakillioglu, T. Grant, S. Velipasalar, and L. Hirshfield, "Classification of fNIRS finger tapping data with multi-labeling and deep learning," *IEEE Sensors J.*, vol. 21, no. 21, pp. 24558–24569, Nov. 2021.

[13] E. Eldele *et al.*, "An attention-based deep learning approach for sleep stage classification with single-channel EEG," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 809–818, 2021.

[14] Z. Wang and T. Oates, "Encoding time series as images for visual inspection and classification using tiled convolutional neural networks," in *Proc. Workshops 29th AAAI Conf. Artif. Intell.*, 2015, pp. 1–7.

[15] A. Dosovitskiy *et al.*, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–22.

[16] W. Wang *et al.*, "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, Sep. 2022.

[17] Z. Liu *et al.*, "Swin transformer: Hierarchical vision transformer using shifted Windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[18] H. Bao, L. Dong, and F. Wei, "BEiT: BERT pre-training of image transformers," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–16.

[19] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2021, pp. 16000–16009.

[20] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollar, and R. Girshick, "Early convolutions help transformers see better," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 30392–30400.

[21] I. Tolstikhin *et al.*, "MLP-mixer: An all-MLP architecture for vision," in *Proc. 35th Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 24261–24272.

[22] H. Touvron *et al.*, "ResMLP: Feedforward networks for image classification with data-efficient training," 2021, *arXiv:2105.03404*.

[23] J.-P. Eckmann, S. O. Kamphorst, and D. Ruelle, "Recurrence plots of dynamical systems," *Europhys. Lett.*, vol. 4, no. 9, pp. 973–977, Nov. 1987.

[24] J. R. Paulo, G. Pires, and U. J. Nunes, "Cross-subject zero calibration driver's drowsiness detection: Exploring spatiotemporal image encoding of EEG signals for convolutional neural network classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 29, pp. 905–915, 2021.

[25] F. Xiao, Y. Chen, and Y. Zhu, "GADF/GASF-HOG: Feature extraction methods for hand movement classification from surface electromyography," *J. Neural Eng.*, vol. 17, no. 4, Jul. 2020, Art. no. 046016.

[26] S. D. Wickramaratne and M. S. Mahmud, "A deep learning based ternary task classification system using gramian angular summation field in fNIRS neuroimaging data," in *Proc. IEEE Int. Conf. E-health Netw., Appl. Services (HEALTHCOM)*, Mar. 2021, pp. 1–4.

[27] A. Vaswani *et al.*, "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[28] H. Liu, Z. Dai, D. So, and Q. V. Le, "Pay attention to MLPs," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 9204–9215.

[29] E. J. Keogh and M. J. Pazzani, "Scaling up dynamic time warping for datamining applications," in *Proc. 6th ACM SIGKDD Intern. Conf. Knowl. Disco. Data Mining*, 2000, pp. 285–289.

[30] D. T. Delpy, M. Cope, P. van der Zee, S. Arridge, S. Wray, and J. Wyatt, "Estimation of optical pathlength through tissue from direct time of flight measurement," *Phys. Med. Biol.*, vol. 33, no. 12, pp. 1433–1442, 1988.

[31] A. Krizhevsky, "Learning multiple layers of features from tiny images," M.S. thesis, Univ. Toronto, Toronto, ON, Canada, 2009.

[32] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *Proc. Int. Conf. Learn. Represent.*, 2019, pp. 1–19.

[33] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *Proc. Int. Conf. Learn. Represent.*, 2017, pp. 1–16.

[34] T. Ishida, I. Yamane, T. Sakai, G. Niu, and M. Sugiyama, "Do we need zero training loss after achieving zero training error?" in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 4554–4564.

[35] J. Shin *et al.*, "Open access dataset for EEG+NIRS single-trial classification," *IEEE Trans. Neural Syst. Rehabil. Eng.*, vol. 25, no. 10, pp. 1735–1745, Oct. 2017.

[36] J. Shin, A. von Lühmann, D.-W. Kim, J. Mehnert, H.-J. Hwang, and K.-R. Müller, "Simultaneous acquisition of EEG and NIRS during cognitive tasks for an open access dataset," *Sci. Data*, vol. 5, no. 1, pp. 1–16, Dec. 2018.

[37] G. Jasdzewski, G. Strangman, J. Wagner, K. K. Kwong, R. A. Poldrack, and D. A. Boas, "Differences in the hemodynamic response to event-related motor and visual paradigms as measured by near-infrared spectroscopy," *NeuroImage*, vol. 20, no. 1, pp. 479–488, Sep. 2003.

[38] B. D. Frederick, L. D. Nickerson, and Y. Tong, "Physiological denoising of BOLD fMRI data using regressor interpolation at progressive time delays (RIPTiDe) processing of concurrent fMRI and near-infrared spectroscopy (NIRS)," *NeuroImage*, vol. 60, no. 3, pp. 1913–1923, Apr. 2012.

[39] B. Lyu *et al.*, "Domain adaptation for robust workload level alignment between sessions and subjects using fNIRS," *J. Biomed. Opt.*, vol. 26, no. 2, pp. 1–21, Jan. 2021.

[40] K. Qing, R. Huang, and K.-S. Hong, "Decoding three different preference levels of consumers using convolutional neural network: A functional near-infrared spectroscopy study," *Frontiers Human Neurosci.*, vol. 14, Jan. 2021, Art. no. 597864.

[41] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 770–778.

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

[43] D. DaJun, L. Jiayan, P. HongXing, and W. Xuebing, "Photoplethysmographic signals identification method based on image coding," in *Proc. Int. Conf. Big Data Economy Inf. Manage. (BDEIM)*, Dec. 2020, pp. 131–136.

[44] D. Dias, U. Dias, N. Menini, R. Lamparelli, G. Le Maire, and R. D. S. Torres, "Image-based time series representations for pixelwise eucalyptus region classification: A comparative study," *IEEE Geosci. Remote Sens. Lett.*, vol. 17, no. 8, pp. 1450–1454, Aug. 2020.

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