Multimodal Isotropic Neural Architecture with Patch Embedding

Hubert Truchan, Evgenii Naumov, Rezaul Abedin, Gregory Palmer, and Zahra Ahmadi$^{(\boxtimes)}$

L3S Research Center, Leibniz University Hannover, Hannover, Germany
{truchan,naumov,abedin,gpalmer,ahmadi}@L3S.de

Abstract. Patch embedding has been a significant advancement in Transformer-based models, particularly the Vision Transformer (ViT), as it enables handling larger image sizes and mitigating the quadratic runtime of self-attention layers in Transformers. Moreover, it allows for capturing global dependencies and relationships between patches, enhancing effective image understanding and analysis. However, it is important to acknowledge that Convolutional Neural Networks (CNNs) continue to excel in scenarios with limited data availability. Their efficiency in terms of memory usage and latency makes them particularly suitable for deployment on edge devices. Expanding upon this, we propose Minape, a novel multimodal isotropic convolutional neural architecture that incorporates patch embedding to both time series and image data for classification purposes. By employing isotropic models, Minape addresses the challenges posed by varying data sizes and complexities of the data. It groups samples based on modality type, creating two-dimensional representations that undergo linear embedding before being processed by a scalable isotropic convolutional network architecture. The outputs of these pathways are merged and fed to a temporal classifier. Experimental results demonstrate that Minape significantly outperforms existing approaches in terms of accuracy while requiring fewer than 1M parameters and occupying less than 12 MB in size. This performance was observed on multimodal benchmark datasets and the authors' newly collected multi-dimensional multimodal dataset, Mudestreda, obtained from real industrial processing devices[1]([1]Link to code and dataset: https://github.com/hubtru/Minape).

Keywords: Multimodal Classification · Isotropic Architecture · Patch Embedding · Time Series

1 Introduction

Humans perceive the world through multiple senses, such as vision and hearing, leading to a multimodal understanding. This combination of data from different modalities offers augmented and complementary information, enabling more robust inference. In recent years, deep learning approaches have made remarkable progress in leveraging data from various modalities, resulting in improved performance across classical problems, including action recognition [14] and semantic

B. Luo et al. (Eds.): ICONIP 2023, LNCS 14447, pp. 173–187, 2024.
https://doi.org/10.1007/978-981-99-8079-6_14

segmentation [1]. Despite these advancements, effectively integrating information from multiple modalities remains a fundamental challenge in multimodal learning. Several research efforts have focused on designing fusion paradigms to fuse multimodal data [12,33]. However, these approaches often require manual design and are specific to certain tasks and modalities, primarily focusing on image and text as the common modalities [31]. Yet, in applications such as intelligent production and healthcare, time series and audio data serve as the major source of information. For instance, device state recognition is a vital problem in intelligent production and healthcare systems, demanding precise monitoring due to the dynamic nature of the device's physical and environmental properties. Maintenance reports show that these devices exhibit symptoms of changing conditions and damage before experiencing complete failure during operation. These symptoms, including distinctive sounds, can serve as indicators for identifying the state of the device. Traditional methods like single modality approaches [22] or self-supervised [18] are impractical in these scenarios due to the complexity and variability of data. Such applications often suffer from noise and conflicts between modalities, which can significantly impact prediction accuracy. Moreover, the limited availability of training samples further complicates the development of efficient models. An ideal algorithm should address these challenges by being robust to noise, selectively leveraging strong modalities, and effectively capturing complementary information among modalities.

In this paper, we present **(Minape)**, a novel multimodal isotropic neural architecture with patch embedding that integrates time series and image data as input. Minape is based on convolutional neural networks (CNNs) and uses a patch-based representation to learn local features from the data. One key advantage of Minape is its isotropic nature, which allows it to handle inputs of varying sizes and aspect ratios and to be trained with fewer labelled examples compared to existing methods. Minape can be employed in numerous applications, including healthcare systems, intelligent production, and real-time monitoring of device states. Through extensive experiments, we demonstrate that Minape achieves significantly higher accuracy compared to the state-of-the-art approaches on both public multimodal classification datasets and our newly introduced multimodal device state dataset (**Mudestreda**) collected from industrial processing devices. In addition, Minape requires significantly less memory to store its much smaller model and can be trained on small and medium size multimodal datasets. Overall, the contributions of this paper are as follows:

- The Minape framework, a novel multimodal isotropic convolutional architecture with patch embedding that requires less than 12 MB size and less than 1M Parameters, which yields the inference output over 90 images/s (results reported for Nvidia 1080TI) for audiovisual data,
- An alternate solution to transformer-based fusion models that can be effectively trained on the small and medium size multimodal datasets with less than 1k instances and capable of operating on edge devices and being deployed in real standalone systems,

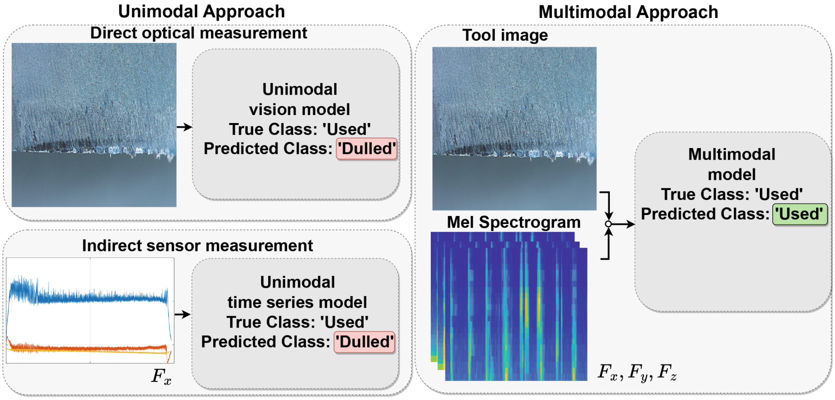


Fig. 1. A sample instance from Mudestreda that shows the importance of leveraging multimodal data for device state prediction.

- Minape yields significantly *higher accuracy* compared to the state-of-the-art approaches. Figure 1 intuitively demonstrates that this augmentation leads to a better predictive performance by an example from Mudestreda, emphasizing its real-world application readiness,
- Minape exhibits *scalability* concerning model pathways and depth dimension. It permits flexible model scaling based on data size and task complexity,
- Finally, we share our collected data, Mudestreda, which is publicly available and aims for multimodal device state recognition.

2 Related Work

2.1 Multimodal Fusion

The focus of multimodal fusion has been primarily on exploring different architectures and techniques for effectively integrating different data modalities to improve the performance of the model. In particular, an Action Segmentation model (ASPnet) based on ASFormer disentangles hidden features into modality-shared components and projects them into independent FC layers [1]. Similarly, a Transformer-based approach [13] creates a single shared representation space using multiple types of image-paired data. [24] also utilized a Transformer-based architecture to leverage audiovisual context in action localization. Other works introduced diverse multimodal fusion strategies involving semantic fusion [31], temporal sequence-based fusion [36], and cross-attention mechanism [18]. Our approach simplifies the fusion process by using isotropic convolutional neural architectures, making it computationally efficient and suitable for smaller datasets.

2.2 Multimodal Model Optimisation

Focusing on model optimization, the primary limitation in the existing literature relates to the resource-demanding nature of the algorithms. In order to manage the long-range temporal dependencies in the data, an attention bottleneck was used, but that solution resulted in high memory demand [1]. Similarly, a joint embedding solution reduced the modalities [13], yet the model is resource-intensive with high memory usage. Another approach proposed an efficient fusion strategy via prompt-based techniques [17] but still required more than 12 GB of memory for training. Other methods also developed innovative ways to optimize their networks, for instance, feature anticipation [36], modality-wise L_2 normalization [32], progressive reduction of tokens [37], but they all required significant computational resources, including large memory and high model size.

Additionally, recent works applied SWIN-transformer [22], TimeSformer [18], VGG-M [27], 3D-Resnet-18 [26] and all reported models with parameters exceeding tens of millions, far above the size of our proposed model. Similarly, recent research showcased innovative techniques with pyramid cross-fusion Transformer [35], contrastive-based alignment training [14], modality dropout training [12], single channel version of ViT [14], knowledge distillation [2], cross-modal prototypical loss [33] and injecting trainable parameters into a frozen ViT [19], but these methods were computationally expensive and required large-scale multimodal datasets for pre-training. Furthermore, recent works presented history-aware [5] and weakly-supervised parsing [20], but these are not feasible for small-to medium-sized multimodal datasets. In contrast, our work proposes a novel, lightweight isotropic convolutional architecture that performs well on limited datasets, requiring less than 1M parameters and less than 12 MB size which is 24M parameter and 88 MB less than the lightest available model.

3 Minape: A Multimodal Isotropic Neural Framework

We consider a target process that generates a sequence of M-modal multi-dimensional data points, $X(1), X(2), \ldots, X(l)$, consisting of time series and images and expressed as the set of tensors $T_{1,1}, T_{1,2}, ..., T_{L,M}$, where $X \in \mathbb{R}^D$ and the m^{th} tensor is $D^{h \times b \times d}_{l,m}$ dimensional in $\mathbb{R}^D$, where $h \times b \times d$ describes tensor dimension, as height, bright and depth respectively, M is the number of auxiliary sensors, and l is the number of data points. Tensors are grouped into g groups according to their modality, where $T_{i,G_1}, i = 1..L, G_1 \in M$ is the set of the tensors belonging to modality type one of the size g_1.

Our objective is to train a classifier $f : \mathbb{R}^D \rightarrow y$ where y is the class of data point $X(i)$. This is measured by sparse categorical cross-entropy in our experiments. We assume the samples are drawn from an ergodic and stationary process. Multiple sensors are grouped in a cluster to reduce the Wasserstein distance between multimodal feature distributions [34].

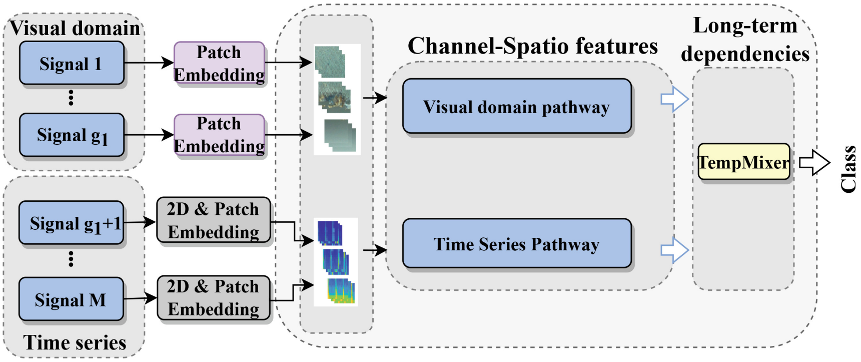


Fig. 2. The Minape general architecture.

To benefit from the synergy of the available M multimodal tensors and enhance the manifolds mapping abilities, we propose deep concatenation of the latent spaces of the linearly embedded two-dimensional representation of input samples from $f_{1:g}$ classifiers to train a classifier $f' : \mathbb{R}^{D_u} \rightarrow y$, where D_u represents the dimension of the deep concatenation layer, and g is the number of modalities of the same type.

3.1 The Minape Architecture Overview

Minape (Fig. 2) is a two-pathway convolutional-based model with visual and time-series feature extractions to obtain a joint representation of the multimodal data. The shared multimodal representation is then passed to the recurrent neural layer to capture the long-term dependencies of the data sequence. The data points comprise M tensors, which are initially grouped into g groups based on their modality type and input into representative $f_i, i = (1 \dots g)$ subnetworks that share parameters across tensors in the same modality group. Then, the learned representations of modalities $u_i, i = (1 \dots g)$ are merged using the concatenation layer, resulting in vector u:

$$u = [f_1(H_1), \dots f_i(H_i) \dots f_g(H_g)], f_i : \mathbb{R}^{D_{H_i}} \rightarrow \mathbb{R}^{D_{u_i}} (i = 1 \dots g), u \in \mathbb{R}^{D_u}. \tag{1}$$

The merged representation, u, serves as the input of the TempMixer block, representing the temporal convolutional network to model the temporal dependencies. This is followed by fully connected and softmax layers:

$$u' = f_{temp}(u),\ \ f_{temp} : \mathbb{R}^{D_u} \rightarrow \mathbb{R}^{D'_u},$$

$$v = W \cdot u' + b, P(y(i)) = \frac{e^{v_i}}{\sum_{j=1}^{y} e^{v_j}}\ for\ i = 1, 2, \dots, y. \tag{2}$$

3.2 Unimodal Feature Representation

The unimodal pathway takes as input data points consisting of g_v tensors T_j, $j = 1 \dots g_v$, of one modality, and stacks them together along their third dimension to create the tensor T_v with $D^{h_v \times b_v \times d_v}$ dimensions. To assure high-precision representation, the tensors are first linearly patched, which decreases their internal resolution, and then fed into a sequence of isotropic channel-point convolutional blocks. The patching procedure and linear embedding are seamlessly integrated into the network [30] using a single convolution layer with d_{g_v} input channels, kernel size k, stride p, and c_{g_v} output channels. This step enables the effective representation and results in the Q_v tensor with the depth of $c_v = g_v \times d_{g_v}$:

$$Q_v = BN(\sigma\{Conv(X_{i,j}, stride = p, kernelSize = k)\}) : D^{h_{g_v}/p \times b_{g_v}/p \times c_v}. \quad (3)$$

The downsampled input is then processed by the Gaussian Error Linear Unit (σ) activation function, which applies nonlinearity by weighting the input by their percentile and can be considered as a smoothed version of the ReLU activation function. The fast version of the GELU [15] is used, which exhibits a sufficient trade-off between the latency and the degree of approximation:

$$H_v = BN(Q_v \cdot \frac{1}{2}[1 + erf(\frac{Q_v}{\sqrt{2}}]). \quad (4)$$

The activation is followed by a batch normalization (BN) layer, which normalizes each channel across mini-batch samples, helping to decrease the susceptibility to variations throughout the data. The channel-point convolutional block consists of the repeated sequence of grouped convolutions and pointwise convolutions with the skip connection that fosters information propagation.

The outcome of the convolution process of each group is combined independently as the dot product of the input and the filters sliding vertically and horizontally across the input field and a sum of the bias term. The grouped convolution with groups equal to the number of channels (c_v) is used. Thus, it allows for a channel-wise separable (depth-wise separable) convolution:

$$H_v^{'} = BN(\sigma\{DepthConv(H_v), groups = c_v\}) + H_v. \quad (5)$$

The depthwise convolution is followed by pointwise convolution, defined as the convolution layer with a kernel size of 1 × 1 and a number of filters equal to the number of channels output by the patch embedding block that allows for linear combinations across channels:

$$H_{v+1} = BN(\sigma\{PointConv(H_v^{'}), kenel = 1 \times 1\}). \quad (6)$$

The last layer is a two-dimensional global average pooling layer that performs downsampling by calculating the average of the vertical and horizontal dimensions of the input volume.

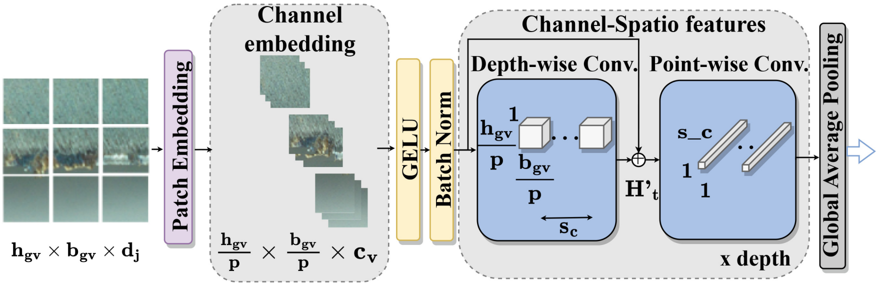


Fig. 3. The structure of the visual modality pathway using patch embeddings.

Visual Domain Pathway: In our architecture, the first pathway groups the tensors in the visual domain into G_v and normalizes their spatial dimensions to meet the subnetwork dimension requirements. These tensors are stacked together along their third dimension to create the tensor T_v with the depth of d_{g_v}:

$$\forall i = 1 \ldots L, \forall j \in G_v \quad norm_v(T_{i,j}) : D^{h_j \times b_j \times d_j} \rightarrow D^{h_{g_v} \times b_{g_v} \times d_{g_v}}. \tag{7}$$

Following the above procedure, as depicted in Fig. 3, we patch and linearly embed the input tensor T_v using a single convolution layer. The outcome of Eq. 3 is then normalized using the fast version of GELU, followed by batch normalization as described in Eq. 4. The resulting tensor H_v serves as input to the sequence of isotropic depth-wise and point-wise convolutions that perform the operations in Eqs. 5 and 6, respectively. The outcome of the visual domain pathway is the vector u_v. The adjustable depth of the isotropic architecture ensures scalability according to the task complexity.

Time Series Pathway: The second pathway (Fig. 4) groups time series tensors into G_s, transforms them into two-dimensional embeddings, normalizes and spatially aligns them to ensure temporal correlations, applies patch embedding, and extracts signal representations with the channel-point convolutions block. The time series sequences of tensor T_j are first resampled to n_{g_s}, and the dimension h_j is trimmed or padded to the h_{g_s} length:

$$\forall i = 1 \ldots L, \forall j \in G_s \quad norm_s(T_{i,j}) : D^{h_j \times b_j \times d_j} \rightarrow D^{h_{g_s} \times b_j}. \tag{8}$$

The two-dimensional embedding is defined as the logarithmically scaled amplitudes of the Mel spectrogram, which performs a windowing transformation of the signal to create a local frequency analysis.

First, the window used for the STFT calculation with a size of $n_{stft} \leq n_{g_s}$ is defined. Due to zero padding and varying amplitudes of T_s, the Hann windowing function is used to normalize the signals. The Mel filter bank is defined as $mel(sr, n_{stft}, n_{mels})$, where sr is the sampling rate of the incoming signal, n_{stft} is the length of the STFT window, and n_{mels} is the number of Mel bands to generate. The output is a two-dimensional array $\mathbb{R}^{n_{mels} \times (1+n_stft/2)}$.

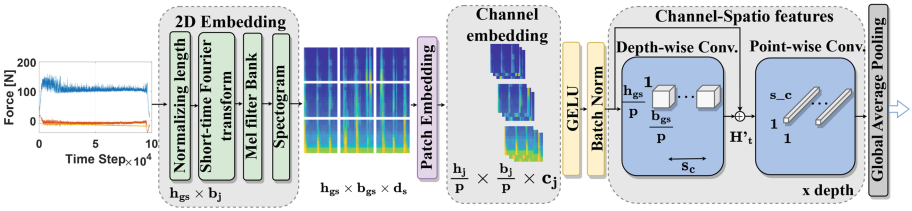


Fig. 4. The structure of the time series modality pathway.

Table 1. Summary of the characteristics of multimodal datasets.

Dataset	Task	Modalities	Types of Modalities	Instances	Classes
Visuo-Haptic [4]	Object recognition	4	Image, Pressure, Texture, Proximity	305	63
MeX [29]	Activity recognition	4	Image, Acceleration, Proximity, Pressure	710	7
BAUM-1a [10]	Emotion recognition	2	Image, Audio	273	9
emoFBVPs [25]	Emotion recognition	2	Image, Audio	1380	23
HA4M [7]	Assembly task	6	RGB images, Depth maps, IR images, RGB-to-Depth-Alignments, Skeleton data	2604	12
Energy [11]	Consumption classification	5	photoplethysmography (PPG), electrocardiography (ECG), Accelerometer, a fraction of oxygen in expired breath (VO2), Gyroscope	1192	3
Mudestreda	Device State recognition	4	Image, Dynamometer data	512	3

The final two-dimensional embedding is obtained using the following formula:

$$2D_{emb} = log_{10}\{mel(STFT(Input_{seq}, stride, hann(n_{stft})))\}, \qquad (9)$$

where $hann(n_{stft})$ refers to the Hanning window function applied with a window length of n_{stft} for the short-time Fourier transformation ($stft$). The stride size determines the sliding window step used for STFT, and the Mel filterbank downsamples the signal to obtain the Mel spectrogram. The obtained two-dimensional representations of tensors grouped in G_s are stacked, resulting in the tensor T_s with $d_s = \sum_{j \in G_s} b_j$ channels:

$$\forall j \in G_s, \forall k = 1 \ldots b_j \quad T_s = 2D_{emb}[T_{:,j}(:,k)] : D^{h_{gs} \times b_{gs} \times d_s}. \qquad (10)$$

Next, the spectrograms are patched and linearly embedded using Eq. 3, followed by the channel-point convolutional block in Eqs. 5 and 6.

4 Experiments

4.1 Experimental Setup

We test the performance of the Minape framework on several open-source multimodal benchmarks (Table 1)[1]. All of them contain visual and time series data: Visuo-Haptic object recognition for robots contains time series (pressure, texture, proximity) and images, MeX human activity recognition contains time series data (acceleration, proximity, pressure) and images, BAUM-1a face mimic recognition and emoFBVPs physiological signals recognition contain images and audio signals, HA4M assembly task recognition contains images (RGB, depth maps, infra-red images, RGB-to-Depth alignments) and time series data (skeleton data), and Energy consumption estimation contains images (photoplethysmography) and time series data (electrocardiography, accelerometer, fraction of oxygen in expired breath, gyroscope).

We present two versions of multimodal fusion neural network architecture: one that uses isotropic neural architecture with patch embedding **Minape**, and the second which uses patch embedding with efficientnetv2-m for intermediate feature representation (**Minape-S**). The TempMixer block consists of four one-dimensional convolution layers with 64 filters each. The model is trained according to an improved training procedure presented in timm [28] that combines best practices for training, such as novel optimization and data augmentation[2]. We have enriched this procedure with the AdamW [23] optimizer and AutoAugment [8]. The results of Minape and Minape-S are reported for depth size $= 16$, 3×10^3 iterations, learning rate $= 0.001$, kernel size $= 13$, embedding dimension $= 1$, *globalAveragePooling* layer, and the GELU activation function.

In addition, we varied the batch size, number of epochs, steps per epoch, the pooling layer (globalMaxPooling2dLayer, globalAveragePooling), and the activation function (GELU, ReLu). We compare Minape with two well-established unimodal algorithms, two recent multimodal versions, and transformers-based architectures. The selected algorithms allow for testing a broad range of fusion methods, e.g., multiplicative combination, embracement, averaging, concatenation, and fusion methods:

- The Temporal Convolutional Network (TCN)[3] leverages dilated causal convolutions and residual blocks for efficiently processing long input sequences and learning complex temporal patterns [3].
- EfficientNetv2-m is a convolution-based image learning method that combines training-aware neural architecture search and scaling, which adaptively adjusts regularization along with image size.
- Multiplicative Multimodal network (Mulmix)[4] uses multiplicative modality mixture combination by first additively creating mixture candidates and then

[1] All experiments were performed on a single Nvidia GTX1080Ti 12 GB GPU.

[2] https://github.com/martinsbruveris/tensorflow-image-models.

[3] https://github.com/locuslab/TCN.

[4] https://github.com/skywaLKer518/MultiplicativeMultimodal.

Table 2. The average accuracy percentage of our proposed framework (Minape-S and Minape with primary learner) compared to other state-of-the-art methods on seven multimodal benchmarks. The results are calculated on five trials and the ranks are reported.

Dataset	TCN*	EfficientNetv2-m**	Mulmix	EmbraceNet	ViT	Minape-S	Minape
Visuo-Haptic	71.9 ± 2.0 (5)	70.2 ± 3.8 (6)	67.6 ± 2.2 (7)	73.3 ± 3.1 (4)	78.1 ± 2.7 (3)	80.7 ± 1.7 (2)	**85.4** ± 2.3 **(1)**
MeX	72.9 ± 3.7 (7)	85.0 ± 4.3 (5)	81.5 ± 3.5 (6)	90.8 ± 3.1 (2)	87.7 ± 3.2 (3)	86.2 ± 4.8 (4)	**93.4** ± 1.9 **(1)**
BAUM-1a	31.3 ± 5.5 (7)	41.8 ± 5.1 (6)	44.0 ± 3.5 (5)	48.2 ± 4.5 (3)	45.6 ± 2.9 (4)	51.5 ± 2.5 (2)	**57.7** ± 3.2 **(1)**
emoFBVPs	81.8 ± 7.2 (7)	83.2 ± 8.0 (6)	86.6 ± 6.0 (4)	83.4 ± 5.8 (5)	89.3 ± 1.8 (2)	88.1 ± 3.3 (3)	**92.7** ± 3.1 **(1)**
HA4M	57.5 ± 9.0 (7)	61.3 ± 6.5 (6)	65.1 ± 4.8 (5)	68.9 ± 6.2 (4)	74.8 ± 1.4 (2)	72.7 ± 3.7 (3)	**76.5** ± 2.9 **(1)**
Energy	65.2 ± 6.1 (7)	67.8 ± 6.6 (6)	72.3 ± 5.7 (5)	74.1 ± 3.9 (4)	77.3 ± 2.7 (3)	78.9 ± 2.1 (2)	**81.6** ± 1.8 **(1)**
Mudestreda	71.4 ± 3.5 (7)	79.9 ± 2.8 (6)	87.1 ± 2.5 (5)	91.3 ± 2.0 (4)	93.7 ± 2.3 (3)	94.7 ± 1.5 (2)	**98.2** ± 1.0 **(1)**
Average rank	6.7	5.9	5.3	3.7	2.9	2.6	1.0

* - trained only on time series, ** - trained only on the visual modality.

selecting useful modality mixtures with multiplicative combination procedure [21].
- EmbraceNet[5] is a deep learning method that leverages cross-modal correlations during the training phases [6]. It uses the *embracement* process to probabilistically select subsets of information from each modality to model the inter-modal correlations.
- Visual Transformer (ViT)[6] uses the Transformer layer with self-attention applied to image patch sequences [9].

4.2 Experimental Results

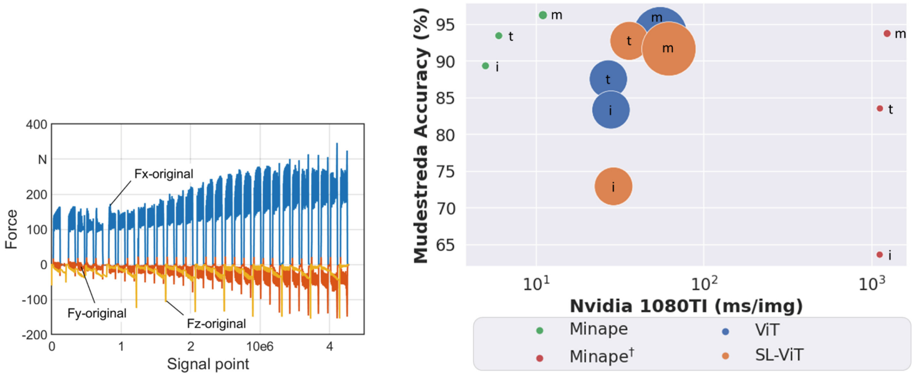
For increased reliability, we performed five independent trials using ten-fold cross-validation by partitioning each dataset into training (80%), validation (10%), and testing (10%) subsets. The outcomes of this comprehensive comparison are detailed in Table 2. For each method, the result of the best hyperparameter setting is reported. All results are in terms of accuracy percentage, and the mean and standard deviation of ten repeats are reported.

The tested datasets have varied sensor sequence lengths ranging from 276 in Visuo-Haptic to 1.5×10^5 in Mudestreda. We change the depth of the Minape architectures from four with a width of 128 in Visuo-Haptic to 16 with a width of 256 in BAUM-1a. We observe that wider networks and successively downsampled convolutional network designs yield better results, even when trained for fewer epochs. Moreover, changing the kernel size from 5 in Mex to 13 in Mudestreda indicates that larger kernels result in better performance. The kernels in the depthwise layer define the size of the receptive field. Thus, large kernels mix arbitrarily distant spatial locations, allowing them to capture spatial dependencies more comprehensively.

Our experiments indicate that the best results belong to patch embedding and isotropic feature extraction architectures. In all tested datasets, Minape

[5] https://github.com/idearibosome/embracenet.

[6] https://github.com/keras-team/keras-io/blob/master/examples/vision/vit_small_ds.py.



(a) (b)

Fig. 5. (a) Example of force signals (Fx, Fy, Fz) from tool nr. 1 (T1) for 30 milling phases. (b) Comparison of multimodal models on Mudestreda. Various modalities are represented with i: image, t: time series, and m: multimodal. Minape† is a model with no patches. Minape (m) is 12 MB and SL-ViT (m) is 863 MB.

achieves significantly better results compared to other methods of comparison. Furthermore, we should note that the corresponding number of subnetworks does not grow exponentially with the number of multimodalities, making the solution scalable, even to high-dimensional spaces. Both the patch embedding and the large kernel size preserve locality well, suggesting that the spatial representation is sensitive to the relative embedding dimensions and filter size and generates equivalent results even without downsampling the representation in subsequent layers. Given that Minape yields better results than Minape-S and the fact that their difference lies only in the convolutional layer architecture, it can be concluded that the performance is more related to the kernel dimension and internal network resolution than to the depth of the classifier.

4.3 Mudestreda: A Multimodal Device State Recognition Use Case

In addition to the existing benchmarks, we study a real-case application of the milling process and propose a new multimodal industrial device state recognition dataset called **Mudestreda**. Mudestreda comprises 512 four-dimensional multimodal observations consisting of three force signal sequences and one RGB image of the shaft milling tool over five weeks from the Production Centrum. The three dimensions of time series modality are forces recorded in three axes, Fx, Fy, and Fz, with a frequency of 10 kHz. After each milling phase, a picture of the tool is taken, and the flank wear is measured to assign each observation to the respective class based on the defined metric: class-1 $[0, 71)\mu m$ (sharp), class-2 $[71, 110)\mu m$ (used), and class-3 $[110, +\infty)\mu m$ (dulled). Figure 5a shows an example of the collected signal, where a strong correlation between the tool wear and the force amplitudes can be observed. It indicates the smallest amplitudes for the sharp tool increase with tool wear.

Table 3. Performance comparison of various models and modalities (image (i), time-series (t), both modalities (m)) on the Mudestreda dataset in terms of accuracy and area under the curve (AUC). Model† results belong to a model without patch embedding. The inference time on the test set and the number of model parameters are reported as ms/img and #Params, respectively. The base model without a primary learner is indicated in **bold**.

Model	Modality	Fusion-Type	ms/img	Size(MB)	#Params	AUC	Accuracy
ViT	i	-	27	421	36.3M	0.89	87.5 ± 3.1
	t	-	28	422	36.4M	0.88	83.3 ± 2.7
	m	Concat	55	844	72.8M	0.98	93.7 ± 2.3
SL-ViT	i	-	29	428	36.5M	0.84	72.9 ± 2.2
	t	-	36	434	36.4M	0.97	92.7 ± 1.8
	m	Concat	62	863	73.7M	0.95	91.6 ± 1.5
Minape	i	-	5	6	0.4M	0.95	**89.3** ± 1.9
	t	-	6	6	0.4M	0.97	**93.4** ± 1.6
	m	**Concat**	**11**	**12**	**0.9M**	**0.98**	**96.2** ± 1.2
Minape †	i	-	1,118	3	0.2M	0.83	63.6 ± 2.6
	t	-	1,122	3	0.2M	0.89	83.5 ± 1.9
	m	Concat	1,239	6	0.4M	0.97	93.7 ± 2.2

Table 3 provides a detailed analysis of the impact of different modalities for Minape and ViT as the second-best comparison method in Table 2. In addition to ViT, we report the results for SL-ViT[7], which is a modified version of the ViT with Shifted Patch Tokenization (SPT) and Locality Self-Attention (LSA) addressing the problem of locality inductive bias and allowing training a small-size datasets [16]. The SL-ViT parameters are set as ViT with flags spt = true, lsa = true.

Table 3 shows that considering the combination of various modalities significantly enhances model accuracy across all examined models. Patches in Minape effectively improve the model's accuracy while reducing inference times. However, transformer models (ViT, SL-ViT), despite their comparable accuracy, have longer inference times and higher memory demands, which limit their applicability. The Minape model stands out with the highest accuracy, lowest memory usage, and shortest latency, making it a top choice for practical applications, given its balanced performance metrics. Figure 5b illustrates that Minape significantly outperforms other models in terms of inference latency and model size (MB). Minape exhibits a model parameter size reduction of more than 80 times compared to the multimodal ViT.

We scrutinized the sensitivity of Minape on the Mudestreda dataset, as shown in Table 4. The results demonstrate that the base model, when utilizing a concatenation fusion strategy, significantly outperforms other fusion types. The key advantage of concatenation fusion is its simplicity, as it does not introduce additional hyperparameters or necessitate complex computations. This not only

[7] https://github.com/keras-team/keras-io/blob/master/examples/vision/vit_small_ds.py.

Table 4. The results for sensitivity analysis of Minape on the Mudestreda dataset. The best model is indicated in **bold**.

Minape	Fusion-Type	Prime Learner	ms/img	Size(MB)	#Param(M)	Accuracy	Δ(%)
Fusion	Add	×	10	6	1.3	87.3 ± 2.8	-8.9
	Multiply	×	12	6	1.3	91.6 ± 1.7	-4.6
	Average	×	11	6	1.3	87.5 ± 2.5	-8.7
Learner	Concat	GRU	10	13	1.9	96.3 ± 1.2	$+0.1$
	Concat	LSTM	10	16	2.1	94.3 ± 1.6	-1.9
	Concat	**TCN**	**11**	**39**	**4.0**	$\mathbf{98.2} \pm 1.0$	$+\mathbf{2}$
Attention ✓	Concat	×	15	8	2.0	95.1 ± 1.4	-1.1
Attention ✓	Concat	TCN	16	17	2.8	94.9 ± 1.8	-1.3

ensures efficient memory utilization but also meets the computational requirements of the system. Among our evaluated primary learners, which include GRU, LSTM, and TCN, the TCN model enhanced the baseline performance by 2% without any impact on the inference time. Our study also underscores the relevance of the attention mechanism in this workflow. Further experiments show that the attention layer does not improve the model performance[8].

5 Conclusion

We proposed Minape, a novel multimodal isotropic neural architecture with patch embedding, to improve multimodal learning for time series and image data. We observed that the patch representation and wider networks that successively down-sample the convolutional network size yield better results than deeper models and even ViT. It uses less memory and has faster inference allowing the use of the model on stand-by devices. The empirical results demonstrate that the modality fusion with the patch embedding yields higher accuracy than state-of-the-art and baseline methods on six multimodal test benchmarks and Mudestreda, our newly introduced real-case multimodal device state recognition dataset.

References

1. van Amsterdam, B., Kadkhodamohammadi, A., Luengo, I., Stoyanov, D.: Aspnet: action segmentation with shared-private representation of multiple data sources. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2384–2393 (2023)
2. Aslam, M.H., Zeeshan, M.O., Pedersoli, M., Koerich, A.L., Bacon, S., Granger, E.: Privileged knowledge distillation for dimensional emotion recognition in the wild. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3337–3346 (2023)

[8] Further ablation studies on the impact of hyperparameters can be found at https://github.com/hubtru/Minape.

3. Bai, S., Kolter, J.Z., Koltun, V.: An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. arXiv:1803.01271 (2018)
4. Bonner, L.E.R., Buhl, D.D., Kristensen, K., Navarro-Guerrero, N.: Au dataset for visuo-haptic object recognition for robots. arXiv preprint arXiv:2112.13761 (2021)
5. Chen, S., Guhur, P.L., Schmid, C., Laptev, I.: History aware multimodal transformer for vision-and-language navigation. Adv. Neural Inform. Process. Syst. (NeurIPS) **34**, 5834–5847 (2021)
6. Choi, J.H., Lee, J.S.: Embracenet: a robust deep learning architecture for multimodal classification. Inform. Fusion **51**, 259–270 (2019)
7. Cicirelli, G., et al.: The ha4m dataset: multi-modal monitoring of an assembly task for human action recognition in manufacturing. Sci. Data **9**(1), 745 (2022)
8. Cubuk, E.D., Zoph, B., Mane, D., Vasudevan, V., Le, Q.V.: Autoaugment: Learning augmentation policies from data. arXiv preprint arXiv:1805.09501 (2018)
9. Dosovitskiy, A., et al.: An image is worth 16x16 words: transformers for image recognition at scale. In: International Conference on Learning Representation (ICLR) (2021)
10. Eroglu Erdem, C., Turan, C., Aydin, Z.: Baum-2: a multilingual audio-visual affective face database. Multimed. Tools Appl. **74**(18), 7429–7459 (2015)
11. Gashi, S., Min, C., Montanari, A., Santini, S., Kawsar, F.: A multidevice and multimodal dataset for human energy expenditure estimation using wearable devices. Sci. Data **9**(1), 537 (2022)
12. Geng, T., Wang, T., Duan, J., Cong, R., Zheng, F.: Dense-localizing audio-visual events in untrimmed videos: A large-scale benchmark and baseline. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22942–22951 (2023)
13. Girdhar, R., et al.: Imagebind: one embedding space to bind them all. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 15180–15190 (2023)
14. Gong, X., et al.: MMG-ego4D: multimodal generalization in egocentric action recognition. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6481–6491 (2023)
15. Hendrycks, D., Gimpel, K.: Gaussian error linear units (gelus). arXiv preprint arXiv:1606.08415 (2016)
16. Lee, S.H., Lee, S., Song, B.C.: Vision transformer for small-size datasets. arXiv preprint arXiv:2112.13492 (2021)
17. Li, Y., Quan, R., Zhu, L., Yang, Y.: Efficient multimodal fusion via interactive prompting. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2604–2613 (2023)
18. Lialin, V., Rawls, S., Chan, D., Ghosh, S., Rumshisky, A., Hamza, W.: Scalable and accurate self-supervised multimodal representation learning without aligned video and text data. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 390–400 (2023)
19. Lin, Y.B., Sung, Y.L., Lei, J., Bansal, M., Bertasius, G.: Vision transformers are parameter-efficient audio-visual learners. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2299–2309 (2023)
20. Lin, Y.B., Tseng, H.Y., Lee, H.Y., Lin, Y.Y., Yang, M.H.: Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing. Adv. Neural Inform. Process. Syst. (NeurIPS) **34**, 11449–11461 (2021)
21. Liu, K., Li, Y., Xu, N., Natarajan, P.: Learn to combine modalities in multimodal deep learning. arXiv preprint arXiv:1805.11730 (2018)

22. Liu, X., Lu, H., Yuan, J., Li, X.: Cat: causal audio transformer for audio classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023)
23. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: International Conference on Learning Representation (ICLR) (2018)
24. Ramazanova, M., Escorcia, V., Caba, F., Zhao, C., Ghanem, B.: Owl (observe, watch, listen): Audiovisual temporal context for localizing actions in egocentric videos. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 4879–4889 (2023)
25. Ranganathan, H., Chakraborty, S., Panchanathan, S.: Multimodal emotion recognition using deep learning architectures. In: IEEE winter conference on Applications of Computer Vision (WACV), pp. 1–9 (2016)
26. Ryan, F., Jiang, H., Shukla, A., Rehg, J.M., Ithapu, V.K.: Egocentric auditory attention localization in conversations. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 14663–14674 (2023)
27. Senocak, A., Kim, J., Oh, T.H., Li, D., Kweon, I.S.: Event-specific audio-visual fusion layers: a simple and new perspective on video understanding. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 2237–2247 (2023)
28. Wightman, R., Touvron, H., Jégou, H.: Resnet strikes back: an improved training procedure in timm. arXiv preprint arXiv:2110.00476 (2021)
29. Wijekoon, A., Wiratunga, N., Cooper, K.: Mex: multi-modal exercises dataset for human activity recognition. arXiv preprint arXiv:1908.08992 (2019)
30. Wu, H., et al.: Cvt: introducing convolutions to vision transformers. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 22–31 (2021)
31. Xiao, Y., Ma, Y., Li, S., Zhou, H., Liao, R., Li, X.: Semanticac: semantics-assisted framework for audio classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023)
32. Xu, R., Feng, R., Zhang, S.X., Hu, D.: Mmcosine: multi-modal cosine loss towards balanced audio-visual fine-grained learning. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023)
33. Xue, Z., Marculescu, R.: Dynamic multimodal fusion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2574–2583 (2023)
34. Zhang, X., Tang, X., Zong, L., Liu, X., Mu, J.: Deep multimodal clustering with cross reconstruction. In: Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD), pp. 305–317 (2020)
35. Zhang, Z., et al.: Abaw5 challenge: a facial affect recognition approach utilizing transformer encoder and audiovisual fusion. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pp. 5724–5733 (2023)
36. Zhong, Z., Schneider, D., Voit, M., Stiefelhagen, R., Beyerer, J.: Anticipative feature fusion transformer for multi-modal action anticipation. In: IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), pp. 6068–6077 (2023)
37. Zhu, W., Omar, M.: Multiscale audio spectrogram transformer for efficient audio classification. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 1–5 (2023)