Exploring Distillation Models for Cultural Heritage Preservation: Traditional Vietnamese Instruments

Thanh Ma, Hieu-Nghia Do, Hieu Nguyen, Ho Doan, Thanh-Nghi Do

Exploring Distillation Models for Cultural Heritage Preservation: Traditional Vietnamese Instruments

Thanh Ma[*,1,3], Hieu-Nghia Do[1], Hieu Nguyen[1], Ho Doan[3], and Thanh-Nghi Do[1,2]

[1] Can Tho University, Vietnam
[2] UMI UMMISCO 209, IRD/UPMC, France
[3] FPT Greenwich Center, FPT University, Can Tho Campus, Can Tho, Vietnam
{mtthanh,dtnghi}@ctu.edu.vn, Hodd@fe.edu.vn
{dhnghia.ctu, hieu10nguyen06}@gmail.com

Abstract. This study investigates the application of distillation models for the preservation of cultural heritage, with a specific emphasis on traditional Vietnamese instruments (TVI). We systematically evaluate various distillation approaches, including combinations of advanced Convolutional Neural Networks (CNNs) and Vision Transformers (ViTs), assessing their performance in terms of model compactness and accuracy. The central aim is to identify a lightweight model that retains high accuracy, ensuring its practical viability in real-world scenarios involving TVI. Our in-depth analysis demonstrates that certain distillation models achieve substantial reductions in computational complexity while preserving the essential classification capabilities crucial for cultural heritage preservation. Notably, the accuracy of these models exceeds 97%, with several combinations reducing model size by approximately 20MB. The detailed evaluation results underscore the potential of these models for efficient management and preservation of heterogeneous cultural datasets.

Keywords: Convolutional Neural Networks · Traditional Vietnamese Instrument · Distillation Models · Cultural Heritage · Vision Transformer.

1 Introduction

Vietnam's cultural heritage is rich and diverse, with traditional musical instruments offering a unique window into its historical and social fabric. These instruments, ranging from the ethereal sounds of the *"Dan Bau"* [4] to the rhythmic beats of the *"cong chieng"* [26], are not only tools for musical expression but also embodiments of Vietnam's artistic ingenuity and cultural identity. Each instrument carries centuries of tradition, meticulously preserved and passed down through generations, reflecting the nation's journey through history. In the realm of computer science, studying and preserving these instruments involves leveraging advanced technologies such as digital sound modeling, machine learning, and virtual reality, enabling a deeper understanding and appreciation of Vietnam's musical legacy. This integration of traditional art forms with modern computational techniques opens new avenues for cultural preservation and dissemination, ensuring that the rich tapestry of Vietnam's musical heritage continues to inspire and educate future generations. In this paper, we pay attention at the problems

of *Traditional Vietnamese Instrument* (TVI for short). Several researches close to this domain include [4,29,16]

From a technological standpoint, artificial intelligence (AI) [15,21] has become an essential tool in preserving traditional instruments by capturing and replicating their unique sounds and playing techniques. Using advanced algorithms and neural networks, AI accurately analyzes and synthesizes the properties and performance styles of these instruments, creating digital replicas that closely mirror their original counterparts. AI's impact spans various domains, notably in intangible cultural preservation, with significant advancements achieved through Convolutional Neural Networks (CNNs) [1,27,20], Distillation Models (DMs) [30], Vision Transformers (ViTs) [10,23], and Diffusion Models [3,2] in computer vision, as well as BERT in natural language processing [11,7] and MFCC + Vector Quantization in audio processing [28,5]. Our research specifically focuses on computer vision, emphasizing CNNs and distillation models for traditional Vietnamese instruments.

Transitioning to CNNs [17], these deep learning models are particularly effective in the detailed analysis and processing of the images of musical instruments. By employing CNNs, we can enhance the accuracy of the models with various architectures (i.e., mobilenet, lenet, and others) instead of utilizing the traditional models (SVM, Decision Tree), ensuring that the nuances and subtleties of each instrument's design are captured with high fidelity. CNNs can identify and extract intricate features from visual data, such as shape, texture, and color, which are crucial for authentically reproducing the appearances of traditional Vietnamese instruments. There are numerous studies related to CNNs in the field of cultural heritage. i.e., [18,19,20,8]. Several pioneering studies have explored the application of CNNs in the preservation and promotion of cultural heritage, both globally and within Vietnam. For example, the study conducted by Marijana et al. [9] explores the classification of architectural heritage images through the application of a deep learning neural network. This research utilized a CNN to sort images into ten distinct categories, including bell towers, stained glass, vaults, and more. The CNN architecture was evaluated on two datasets: one encompassing all ten categories and another narrowed down to five categories. The results revealed impressive accuracy rates, achieving up to 90% for both dataset configurations, underscoring the efficacy of CNNs in the classification of cultural heritage images. Recent advancements in computer vision and machine learning have led to innovative methods for evaluating traditional costumes such as the Vietnamese *Aodai*. For instance, Pham et al. [6] applied CNN-based techniques to develop a sophisticated recognition system, thereby facilitating the digital preservation and systematic cataloging of these culturally significant garments.

Our objective is to develop a model optimized for deployment on systems with limited hardware capabilities, such as smartphones and Arduino devices, while ensuring real-time operation. Recognizing that many CNN models are substantial in size with extensive parameters, we focus on refining these models to preserve traditional musical instruments on compact media platforms. Our vision is to engage younger audiences, particularly children, who often show limited interest in these cultural artifacts. The ultimate goal is to create a system capable of identifying musical instruments directly through a mobile phone camera with standard configuration, providing users with rele-

vant information about the recognized instruments. This research lays a crucial foundation for the development of such a system, with particular emphasis on reducing model size to meet the constraints of mobile platforms.

Distillation models [12,24] offer an effective solution by balancing efficiency and accuracy, compressing complex neural networks into more deployable forms with minimal performance loss. This approach is crucial for preserving the visual integrity of traditional musical instruments, ensuring both authenticity and accessibility. By leveraging distillation, we create precise digital representations of Vietnam's traditional instruments, enhancing their integration into educational programs, virtual reality, and digital archives. Our key contribution includes developing a self-curated image dataset and exploring various CNN combinations within distillation frameworks, focusing on optimized image classification. Specifically, we tackle this problem through distillation models. We conduct extensive experiments with a diverse array of CNN architectures to provide a vivid and comprehensive perspective on machine learning models tailored for the cultural preservation of Vietnamese Traditional Instruments (VTIs). As mentioned above, our approach aims to lay the groundwork for traditional instrument recognition systems and heralds the development of our ViTIP System *(Vietnamese Traditional Integument Preservation System)* [25]. Note that, this paper will not detail the ViTIP system itself; instead, we will delve into the exploration of CNN models to identify the optimal combinations that ensure *lightweight models while maintaining high accuracy.*

The structure of this paper is as follows: Section 2 offers a brief overview of the fundamental background. In Section 3, we introduce the proposed approaches. Section 4 details the experimental setup and analyzes the results of the summary models. Finally, Section 5 concludes the paper and outlines potential directions for future research.

2 Background

In this section, we will present a comprehensive overview of TVIs, accompanied by an in-depth exploration of two advanced AI techniques: Convolutional Neural Networks and Distillation Models.

2.1 Traditional Vietnamese Instrument

Traditional Vietnamese instruments form a crucial component of Vietnam's rich and diverse musical heritage, deeply interwoven with the nation's cultural and historical tapestry. These instruments, employed across various traditional music genres such as folk music, court music, and religious ceremonies, showcase the unique sounds and techniques that have been passed down through generations. One of the most iconic traditional instruments is the "dan_bau", a monochord zither known for its ethereal and haunting tones. This instrument is often used in solo performances and traditional Vietnamese orchestras. Another significant instrument is the "dan_tranh", a 16-string zither that is similar to the Chinese *guzheng*. It is prized for its versatility and ability to produce a wide range of sounds, from delicate and intricate melodies to powerful and resonant tones. The "dan_nguyet", or moon lute, is a two-stringed lute with a round body, known for its deep, mellow sound, often used in traditional folk music and hát

văn, a form of spiritual singing. The "*dan_ty_ba*", a pear-shaped lute, is used in both folk and classical music and is admired for its expressiveness and the complexity of its playing techniques. Percussion instruments also play a vital role in traditional Vietnamese music. The "*trong*" (drum) family includes various sizes and types, each with its own distinct sound and purpose. These drums are integral to many musical ensembles and ceremonial events.

Fig. 1: Various Traditional Instrument Types in Vietnam.

Several traditional Vietnamese music forms and their associated instruments have been recognized by UNESCO as Intangible Cultural Heritage. "*Nha_nhac_cung_dinh_Hue*" (Hue royal court music)[4] is one such example, featuring instruments like the *dan_ty_ba, dan_nguyet*, and a variety of percussion instruments. This form of music, performed during the Nguyen Dynasty, represents the pinnacle of Vietnamese court music with its refined and sophisticated compositions. Another notable example is "*ca_tru*" [22], an ancient genre of chamber music that employs the *danday* (a long-necked three-stringed lute) and "*phach*" (small wooden sticks beaten on a bamboo block). *ca_tru* performances are intimate and feature intricate vocal techniques, rhythmic complexity, and poetic lyrics. UNESCO's recognition of *ca_tru* highlights its cultural importance and the need for its preservation.

These instruments and the music they produce not only represent Vietnam's artistic achievements but also serve as a testament to the country's rich cultural identity. The recognition by UNESCO underscores the global significance of these traditions and the importance of safeguarding them for future generations.

[4] https://ich.unesco.org/en/RL/nha-nhac-vietnamese-court-music-00074

2.2 Convolutional Neural Network

Convolutional Neural Networks (CNNs) [1,27] are a pivotal deep learning architecture, specifically designed for processing structured grid data like images. CNNs utilize a hierarchical structure with layers such as convolutional, pooling, and fully connected layers. Convolutional layers apply filters to extract complex features, while pooling layers reduce dimensionality, enhancing efficiency without losing critical information. Fully connected layers, positioned at the network's end, synthesize these features for high-level reasoning and classification. CNNs' adaptive learning of spatial hierarchies makes them essential in domains like image and video recognition, natural language processing, and medical image analysis. The general CNN Model is presented in Figure 2. The core architecture of a CNN comprises several key components:

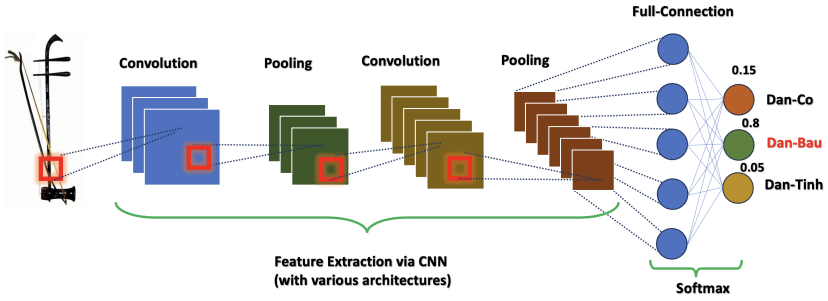

Fig. 2: the CNN Architecture.

- **Input Layer:** This layer accepts the raw pixel values of the input image, typically in the form of a three-dimensional matrix (height, width, and channels).
- **Convolutional Layers**: These layers apply a series of convolutional filters (kernels) to the input image or the output of the previous layer. Each filter scans the input and produces a feature map that highlights specific features such as edges, textures, or colors. The operation can be represented as:

$$X \times W(i,j) = \sum \sum X(i+m, j+n) \times W(m,n)$$

  where X is the input, W is the filter, n and m are indices for the position within the filter and (i,j) represents the spatial coordinates.
- **Activation Function (ReLU)**: After convolution, an activation function, typically the Rectified Linear Unit (ReLU), is applied to introduce non-linearity into the model:

$$f(x) = \max(0, x)$$

- **Pooling Layers**: These layers reduce the spatial dimensions of the feature maps, thereby decreasing the computational load and the number of parameters. The most

common form is max pooling, defined as:

$$P(i, j) = \max_{m,n} X(i + m, j + n)$$

where P is the pooled output, and m, n define the pooling window.

– **Fully Connected Layers**: These layers are akin to the traditional neural networks, where each neuron is connected to every neuron in the previous layer. They integrate the high-level features learned by the convolutional and pooling layers to classify the input into various categories. The operation can be formulated as:

$$y = f(W \times x + b)$$

where W is the weight matrix, x is the input vector, b is the bias, and f is the above activation function.

– **Output Layer**: This layer provides the final classification output, typically through a softmax activation function for multi-class classification tasks:

$$\Delta(v_i) = \frac{\exp^{v_i}}{\sum_j \exp^{v_j}}$$

where v_i is the $i - th$ element of the input vector to the softmax function

In this paper, we expect to investigate ten different architectures, specifically *MobileNet, MobileNetV2, DenseNet201, DenseNet121, InceptionV3, Xception, NASNet-Mobile, ResNet50, VGG19, and VGG16* and one vision transformer. We select these architectures based on the following criteria: (1) Diversity in architectural sizes (ranging from lightweight to heavyweight models); (2) Inclusion of both recent and older architectures; (3) Richness in the implementation of activation functions. These criteria are intended to enhance the integration into distillation models, thereby making the process more "colorful and promising". Next, we will delve into the distillation model, which serves as the focal fundamental of this paper.

2.3 Distillation Models

Distillation models [12,24,14] (DMs for short) represent a significant advancement in the field of machine learning, particularly in the domain of model compression and efficiency. These models aim to transfer knowledge from a large, complex model (often referred to as the "teacher") to a smaller, more efficient model (the "student"), without significant loss of performance. The process involves training the student model to mimic the behavior and output of the teacher model, often through techniques such as *soft target training and intermediate feature matching*. This not only reduces the computational resources required for deployment but also enables the deployment of sophisticated models in resource-constrained environments, such as mobile devices or edge computing scenarios. The efficacy of DMs lies in their ability to retain critical information and generalize well from the teacher model, making them a powerful tool for optimizing machine learning applications.

DMs offer several advantages over traditional CNN [13]. They are as follows: (1) Firstly, they significantly reduce the computational resources required for training and

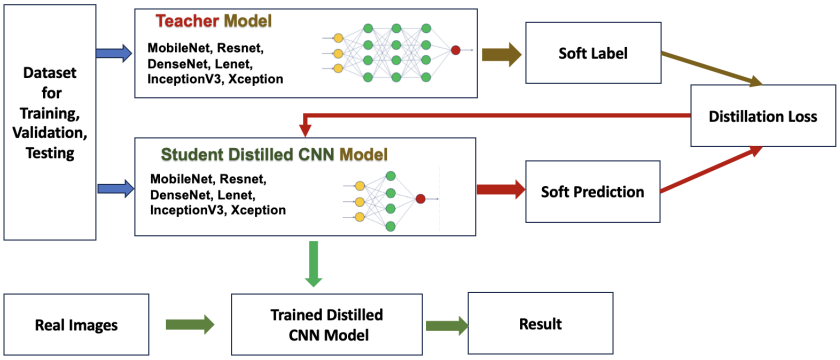

Fig. 3: Framework of the distillation model.

inference, making them ideal for deployment on devices with limited processing power, such as mobile phones and embedded systems. This efficiency is achieved by compressing knowledge from a large, complex teacher model into a smaller, more manageable student model without significant loss of accuracy. (2) DMs often exhibit enhanced generalization capabilities, capturing more nuanced patterns in the data and improving performance on unseen test sets. This is particularly beneficial for applications where robustness and reliability are crucial. (3) DMs facilitate the transfer of knowledge between different architectures and domains. They can leverage pre-trained teacher models to rapidly adapt to new scenarios, accelerating the development cycle and expanding the applicability of machine learning solutions. The framework of the distillation models is presented in Figure 3.

The core idea of DM is to train the student model to match the output distribution of the teacher model. This is achieved through a loss function that combines the traditional classification loss with a distillation loss, which measures the divergence between the teacher's and student's output probabilities. Let $\mathcal{T}$ *be the temperature* parameter that smooths the output probabilities of the teacher model. The higher the temperature, the softer the probability distribution.

- **Soft Targets** denoted by ϕ: The teacher model's softened outputs are given by:

$$\phi_i = \frac{\exp(\frac{z_i}{\mathcal{T}})}{\sum_j \frac{\exp(z_j)}{\mathcal{T}}}$$

  where z_i are the logits (pre-softmax activations) of the teacher model.
- **Student Model Output**: Let $\mathcal{S}$ denote the student model's output probabilities at temperature $\mathcal{T}$.
- **Distillation Loss**: The *Kullback-Leibler* (KL) divergence between the softened outputs of the teacher and student models is defined as:

$$\mathcal{L}_{Distillation} = \mathcal{T}^2 \times \sum_i \phi_i \times \log(\frac{\phi_i}{\mathcal{S}})$$

– **Classification Loss**: The standard cross-entropy loss on the hard labels is given by:

$$\mathscr{L}_{Student} = \mathscr{L}_{CE} = -\sum_{i} y_i \times \log(\mathcal{S})$$

where y_i are the true labels.

– **Total Loss:** The overall loss function combines both components:

$$\mathscr{L}_{final} = \alpha \times \mathscr{L}_{Student} + (1 - \alpha) \times \mathscr{L}_{Distillation}$$

where α is a weighting factor that balances the contribution of the classification and distillation losses.

It is important to emphasize that we assess multiple distillation models with the aim of striking an optimal balance between a streamlined architecture and preserving robust accuracy.

3 Our Approaches and Strategies

The paper employs various approaches and strategies to compare model sizes. These strategies are detailed as follows:

1. **Lightweight Teacher, Lightweight Student Approach**: This approach leverages a compact teacher model to transfer knowledge into an equally compact student model. The objective is to rigorously evaluate the efficacy of the student model in capturing critical information from a "smaller" teacher while still achieving robust performance metrics. This methodology seeks to determine whether minimizing model size can still preserve essential learning outcomes.
2. **Heavyweight Teacher, Lightweight Student Approach**: In this strategy, a large or complex teacher model is utilized to distill knowledge into a smaller student model. The focus here is on understanding the extent to which a lightweight student can assimilate the sophisticated knowledge of a more advanced teacher. This approach is crucial for evaluating the trade-offs between the complexity of the model and its resulting performance, providing insights into how much simplification is possible without significant loss in accuracy.
3. **Lightweight Teacher, Heavyweight Student Paradigm:** This approach involves deploying a lightweight teacher model to impart knowledge to a more complex, larger student model. The objective is to explore whether a streamlined teacher can effectively enhance the learning process of a more sophisticated student, potentially elevating the overall capabilities of the model.
4. **CNN Architecture Alignment Between Teacher and Student:** In this methodology, both the teacher and student models are designed with the same CNN architectures. By standardizing the architecture while varying only the model sizes, the study rigorously examines the influence of model size on performance, thereby isolating and controlling for architectural variances.

5. **Investigation of Vision Transformer (ViT) Teacher and CNN Student:** This paper conducts an in-depth analysis of the knowledge distillation process between a ViT-based teacher model, renowned for its proficiency in image classification tasks through attention mechanisms, and a CNN-based student model. The objective is to rigorously evaluate the compatibility and effectiveness of knowledge transfer between these distinct neural network architectures, thereby providing insights into the potential synergies and limitations inherent in cross-architecture distillation.

Each strategy allows for a nuanced comparison of model sizes and architectures, offering insights into optimal configurations for various computational and performance requirements in practical applications. Now, we will present our experimental outcomes.

4 Experimental Results

This section details the collected dataset and provides an empirical comparison of the classification models. Our dataset and implementation is published in Github[5]

4.1 Dataset and Implementation Environment

We collect a diverse dataset of musical instrument images sourced from various channels, including video extractions, online repositories, and carefully curated manual collections. Some instruments, such as the Dan_Co and $Trong\_Quan$, posed challenges during the acquisition process. Despite this, our primary aim is to evaluate the model's effectiveness in identifying these instruments. As a result, the dataset contains varying quantities of images with significant distributional differences. For example, the Dan Co set includes 193 images, while the Dan_Tranh set comprises 858 images. Further details on these instruments are provided in Table 1. For training and testing, we split the dataset with an 80 : 20 ratio.

To evaluate our approach and the training model, we utilize the AI libraries *(i.e., TensorFlow's (version 2.16.1), Keras (version 3.4.1))* and run the experiments on the computer with the following configuration: AMD Ryzen 7 6800HS, Creator Edition 3.20 GHZ, 16 GB RAM, Windows 10 OS.

4.2 Results of CNNs

As outlined earlier, we conducted evaluations on 10 distinct CNN architectures. Prior to presenting the outcomes of the distillation models, we intend to demonstrate the performance of each CNN model specifically on the TVIs dataset. We offer insights into Accuracy, Precision, Recall, and F1 metrics under the same parameter set: Input size set at 224×224×3, a batch size of 32, and employing early stopping with a patience level of 5. Moreover, we are eager to explore Vision Transformers (ViTs) models in our research endeavors.

Intuitively, CNN models demonstrated excellent performance, with most achieving over 98% accuracy. However, the Resnet50 architecture was less suitable for this

[5] https://github.com/dohieunghia2002/DistillationModels-ViTIP.git

Table 1: Collected Dataset for TVIs.

ID	Instrument Name	Number of VTI images	Note/English Name
1	*Cong_Chieng*	271	Gong
2	*Dan_Bau*	360	One-string zither
3	*Dan_Co*	193	2-chord fiddle
4	*Dan_Da*	598	Lithophone
5	*Dan_Day*	425	-
6	*Dan_Nguyet*	679	Moon Lute
7	*Dan_Sen*	707	-
8	*Dan_T'rung*	800	-
9	*Dan_Tinh*	636	Gourd lute
10	*Dan_Tranh*	858	-
11	*Dan_Ty_Ba*	278	Pipa
12	*Khen*	261	-
13	*Trong_Quan*	215	military drum
		6,281	

particular image dataset, yielding only about 85% accuracy. In terms of model size, lightweight architectures like MobileNet and MobileNetV2 are approximately 24 MB, while some more complex architectures, such as Resnet450, Xception, and InceptionV3, resulted in models exceeding 100 MB. Moreover, we also evaluated the dataset using the Vision Transformer (ViT) model. Given that the ViT model is derived from the Transformer architecture, it is inherently suited for large learning models (LLMs), which results in its substantial size of over 225 MB. The model achieved an accuracy of approximately 97.5%. However, this accuracy is lower than that of CNN architectures, as the ViT model necessitates a sufficiently large and balanced dataset, rather than a noisy one, to perform optimally. Now, we present the results of the DM model in the following section.

4.3 Results of Distillation Models

To comprehensively evaluate these models, we adopted three distinct strategies: a lightweight student model, a lightweight teacher model, and a heavyweight student model, all of which share the same architecture. Additionally, we present results where the Teacher model utilizes the ViT architecture and the Student model employs the CNN architecture. The evaluation outcomes are detailed in Table 3.

Our exploration into distillation models encompassed a diverse array of 29 different methodologies (see Figure 3). Notably, our findings underscored a crucial interplay between model size and accuracy when comparing heavy teacher models with lightweight student counterparts. While the physical footprint of the model largely hinges on the student's architecture, the fidelity of predictions heavily leans on the teacher's prowess.

Interestingly, in several instances, leveraging a robust teacher model facilitated superior performance in the student model compared to traditional CNN approaches. Naturally, we observed a nuanced trade-off: while the accuracy from teacher to stu-

Table 2: Results of the instrument classification with CNNs

ID	Models	Size (MB)	Accuracy (%)	Precision (%)	Recall (%)	F1 (%)
1	CNN(ResNet50)	114.71	85.11	86.61	85.11	84.32
2	CNN(VGG16)	62.4	97.56	97.59	97.56	97.56
3	CNN(VGG19)	82.64	97.08	97.13	97.08	97.08
4	CNN(Mobilenet)	24.76	99.21	99.27	99.18	99.21
5	CNN(MobilenetV2)	24.28	97.48	97.5	97.48	97.48
6	CNN(DenseNet201)	94.67	98.34	98.33	98.42	98.36
7	CNN(DenseNet121)	40.29	98.37	98.33	98.39	98.37
8	CNN(InceptionV3)	108.27	98.81	98.80	98.83	98.81
9	CNN(Xception)	104.18	98.56	98.57	98.55	98.56
10	CNN(NASNetMobile)	31.17	97.08	97.1	97.08	97.09
11	Vision Transformer	225.16	97.57	97.57	97.61	97.58

dent typically saw a modest decline, this was accompanied by negligible deviations, often less than 1% in accuracy. Moreover, our experiments with configurations featuring lightweight teachers and heavier students generally yielded marginal improvements in accuracy, accompanied by modest reductions in model size. Conversely, aligning both teacher and student models structurally led to significant size reductions, often slashing model sizes by $7MB$ or more

4.4 Comparison and Discussion

When comparing CNNs and DM, a few key observations stand out. First and foremost, the DM models tend to be significantly lighter than the traditional CNN models. This reduction in model size does not come at the cost of performance, as the accuracy of DM models remains nearly on par with CNNs, exhibiting only a minor deviation of $0.2 - 1.0\%$, which is considered negligible in most practical scenarios.

One particularly noteworthy finding is the superior performance achieved by combining the Vision Transformer (ViT) with ResNet50. This hybrid approach not only surpasses the performance of ResNet50 alone but also benefits from a reduced model size. The synergy between the advanced representation capabilities of ViT and the robust feature extraction of ResNet50 results in a model that is both more efficient and more accurate, highlighting the potential of integrating diverse architectural paradigms within distillation frameworks. In this study, we do not explore Vision Transformers (ViT) as student models due to their large number of parameters, which limits their ability to significantly reduce model size and meet real-time processing requirements.

In the context of our study, model distillation demonstrates significant advantages in terms of model size and efficiency. By transferring knowledge from a larger, more complex teacher model to a smaller student model, we achieve comparable performance while significantly reducing computational requirements and memory footprint. This makes the student model particularly suitable for deployment in resource-constrained environments such as mobile devices and embedded systems.

Table 3: Results of the instrument classification with Distilation Models.

ID	Teacher Model	Student Model	Size (MB)	Acc (%)	Precision (%)	Recall (%)	F1 (%)
1	VGG16	NASNetMobile	22.8	96.97	97.24	96.97	96.99
2	VGG19	NASNetMobile	22.8	97.92	97.95	97.92	97.92
3	DenseNet201	NASNetMobile	22.8	96.17	96.59	96.17	97.23
4	DenseNet121	NASNetMobile	22.8	97.13	97.19	97.13	97.15
5	InceptionV3	NASNetMobile	22.8	97.37	97.47	97.137	97.39
6	Xception	NASNetMobile	22.8	98.03	98.13	98.08	98.11
7	VGG16	MobileNetV2	14.16	96.73	96.95	96.73	96.64
8	VGG19	MobileNetV2	14.16	98.16	98.26	98.16	98.17
9	DenseNet201	MobileNetV2	14.16	97.45	97.52	97.45	97.47
10	DenseNet121	MobileNetV2	14.16	98.64	98.67	98.64	98.65
11	InceptionV3	MobileNetV2	14.16	96.73	96.91	96.73	96.63
12	Xception	MobileNetV2	14.16	95.93	96.76	95.93	96.08
13	VGG16	MobileNet	16.65	99.36	99.36	99.39	99.38
14	VGG19	MobileNet	16.65	99.04	99.07	99.04	99.05
15	DenseNet201	MobileNet	16.65	99.12	99.14	99.12	97.12
16	DenseNet121	MobileNet	16.65	99.28	99.24	99.32	99.28
17	InceptionV3	MobileNet	16.65	98.80	96.83	98.8	98.81
18	Xception	MobileNet	16.65	97.05	97.53	97.05	97.09
19	Mobilenet	VGG16	58.26	97.61	97.69	97.61	97.66
20	Mobilenet	InceptionV3	92.15	98.88	98.91	98.88	98.88
21	Mobilenet	DenseNet201	79.55	99.2	99.30	99.15	99.22
22	DenseNet121	InceptionV3	92.15	98.72	98.78	98.72	98.75
23	InceptionV3	DenseNet201	79.55	98.32	96.40	98.32	98.33
24	Xception	DenseNet201	79.55	98.56	98.60	98.56	98.59
25	MobileNet	MobileNet	16.65	95.46	95.64	95.46	95.52
26	NASNetMobile	NASNetMobile	22.8	97.85	97.95	97.85	97.89
27	InceptionV3	InceptionV3	93.25	98.00	98.14	98.04	98.08
28	Xception	Xception	88.07	98.48	98.56	98.48	98.50
29	Vision Transformer	Resnet50	95.87	95.14	98.14	95.40	98.04

However, this approach is not without its drawbacks. One major limitation is the potential loss of accuracy during the distillation process. Despite the student model mimicking the teacher model's behavior, it may not fully capture all the nuances and intricacies of the teacher model, leading to a degradation in performance. Additionally, the process of distillation itself can be computationally intensive and time-consuming, as it involves training two models and ensuring effective knowledge transfer. Moreover, there is a risk of overfitting to the teacher model's specific outputs rather than generalizing well to unseen data. This can result in a student model that performs well on tasks similar to those seen during distillation but struggles with more diverse or novel inputs. Another concern is the sensitivity of the distillation process to hyperparameter settings, such as the temperature parameter used to soften the teacher's predictions and the balance between the distillation loss and the original task loss. Incorrect tuning of these parameters can further exacerbate performance issues. Furthermore, the initial requirement of a highly performant teacher model poses an additional challenge. Developing such a model can be resource-intensive and may not always be feasible, especially in scenarios where computational resources or time are limited. The dependency on the quality of the teacher model means that any deficiencies in the teacher will inevitably

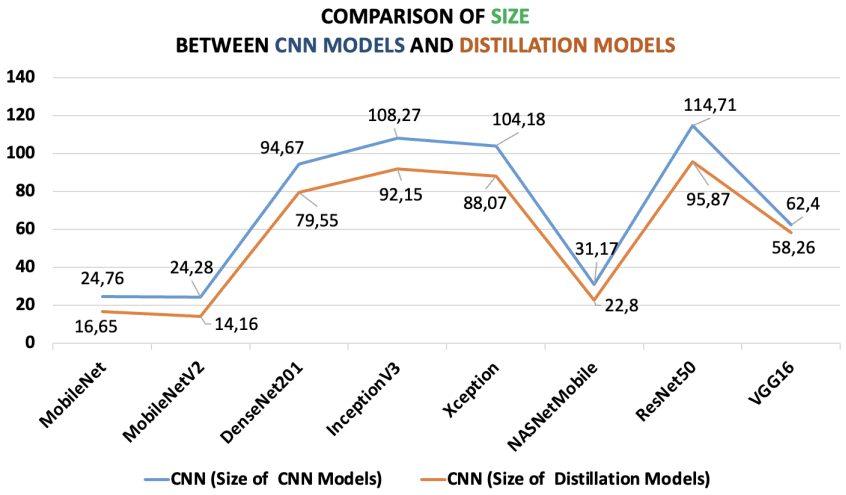

Fig. 4: Size Comparison of CNN vs Distillation Models.

be transferred to the student. The comparasion of CNN and Distillation models is presented in Figure 4.

5 Conclusion and Future Works

In conclusion, this study presented an in-depth exploration of distillation models applied to the preservation of cultural heritage, focusing on traditional Vietnamese musical instruments. This research serves as one of the crucial foundations in the series of studies aimed at developing the ViTIPs system (as mentioned in the introduction). Our primary contribution lies in the creation and utilization of a comprehensive dataset of traditional Vietnamese instruments, which was meticulously collected to ensure its authenticity and richness. By comparing various distillation models with the baseline CNN, we were able to demonstrate the effectiveness of these models in achieving a desirable balance between model compactness and accuracy. Our results indicate that while some lightweight distillation models can indeed match the performance of the traditional CNN in terms of accuracy, they offer the added benefit of reduced computational complexity and resource requirements. This highlights the potential of these models in practical applications where computational efficiency is paramount.

Looking forward, future works will focus on several key areas. First, we aim to expand our dataset to include more diverse instruments and additional metadata, thereby enhancing the robustness and generalizability of our models. Second, we plan to explore advanced distillation techniques, such as multi-stage and ensemble distillation, to further improve model performance. Lastly, we intend to investigate the deployment of these lightweight models in real-world applications (ViTIP system), such as mobile and

edge computing environments, to facilitate wider access to and preservation of cultural heritage artifacts.

Acknowledgements: This study is funded by the Can Tho University, Code: T2024-86. Moreover, Thanh MA has also received support from the European Union's Horizon research and innovation program under the MSCA-SE (Marie Skłodowska-Curie Actions Staff Exchange) grant agreement 101086252; Call: HORIZON-MSCA-2021-SE-01; Project title: STARWARS (STormwAteR and WastewAteR networkS heterogeneous data AI-driven management).

References

1. Abed, M.H., Al-Asfoor, M., Hussain, Z.M.: Architectural heritage images classification using deep learning with cnn. CEUR-WS (2020)
2. Amadeus, M., Castañeda, W.A.C., Zanella, A.F., Mahlow, F.R.P.: From pampas to pixels: Fine-tuning diffusion models for ga\'ucho heritage. arXiv preprint arXiv:2401.05520 (2024)
3. An, L., Zhou, P., Zhou, M., Wang, Y., Geng, G.: Diffusion transformer for point cloud registration: digital modeling of cultural heritage. Heritage Science **12**(1), 198 (2024)
4. Beebe, L.: The Vietnamese Dàn Bâu: A Cultural History of an Instrument in Diaspora. University of California, Santa Cruz (2017)
5. Bhatt, M., Patalia, T.: Neural network based indian folk dance song classification using mfcc and lpc. International Journal of Intelligent Engineering & Systems **10**(3) (2017)
6. Cao, T., Nguyen, H.T., Nguyen, H.M., Hoshino, Y.: Modeling emotional evaluation of traditional vietnamese aodai clothes based on computer vision and machine learning. Industrial Applications of Affective Engineering pp. 111–122 (2014)
7. Catelli, R., Bevilacqua, L., Mariniello, N., Di Carlo, V.S., Magaldi, M., Fujita, H., De Pietro, G., Esposito, M.: A new italian cultural heritage data set: detecting fake reviews with bert and electra leveraging the sentiment. IEEE Access **11**, 52214–52225 (2023)
8. Chau, N.K., Ma, T.T., et al.: An automatic detection of fundamental postures in vietnamese traditional dances. In: CS & IT Conference Proceedings. CS & IT Conference Proceedings (2020)
9. Ćosović, M., Janković, R.: Cnn classification of the cultural heritage images. In: 2020 19th International Symposium INFOTEH-JAHORINA (INFOTEH). pp. 1–6. IEEE (2020)
10. Duan, X., Jiang, C., Fan, Y.: Enhanced inpainting model revitalizes historical paintings with vision transformer. In: 2023 9th International Conference on Virtual Reality (ICVR). pp. 582–589. IEEE (2023)
11. Farella, M., Chiazzese, G., Bosco, G.L.: Question answering with bert: designing a 3d virtual avatar for cultural heritage exploration. In: 2022 IEEE 21st Mediterranean Electrotechnical Conference (MELECON). pp. 770–774. IEEE (2022)
12. Gou, J., Yu, B., Maybank, S.J., Tao, D.: Knowledge distillation: A survey. International Journal of Computer Vision **129**(6), 1789–1819 (2021)
13. Hong, Y.W., Leu, J.S., Faisal, M., Prakosa, S.W.: Analysis of model compression using knowledge distillation. IEEE Access **10**, 85095–85105 (2022)
14. Huang, T., You, S., Wang, F., Qian, C., Xu, C.: Knowledge distillation from a stronger teacher. Advances in Neural Information Processing Systems **35**, 33716–33727 (2022)
15. Hunt, E.B.: Artificial intelligence. Academic Press (2014)
16. Jährlichen, G.: Uniqueness re-examined: The vietnamese lute (dan day). Yearbook for traditional music **43**, 147–179 (2011)

17. Li, Z., Liu, F., Yang, W., Peng, S., Zhou, J.: A survey of convolutional neural networks: analysis, applications, and prospects. IEEE transactions on neural networks and learning systems **33**(12), 6999–7019 (2021)
18. Liu, E.: Research on image recognition of intangible cultural heritage based on cnn and wireless network. EURASIP Journal on Wireless Communications and Networking **2020**(1), 240 (2020)
19. Lu, Y., Zhou, J., Wang, J., Chen, J., Smith, K., Wilder, C., Wang, S.: Curve-structure segmentation from depth maps: A cnn-based approach and its application to exploring cultural heritage objects. In: Proceedings of the AAAI Conference on Artificial Intelligence (2018)
20. Ma, T.T., Benferhat, S., Bouraoui, Z., Tabia, K., Do, T.N., Pham, N.K.: An automatic extraction tool for ethnic vietnamese thai dances concepts. In: 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA). pp. 1527–1530. IEEE (2019)
21. Negnevitsky, M.: Artificial intelligence: a guide to intelligent systems. Pearson education (2005)
22. Norton, B.: Music revival, ca trù ontologies, and intangible cultural heritage in vietnam. In: The Oxford handbook of music revival, pp. 158–179. Oxford University Press New York (2014)
23. Réby, K., Guilhelm, A., De Luca, L.: Semantic segmentation using foundation models for cultural heritage: an experimental study on notre-dame de paris. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 1689–1697 (2023)
24. Ruffy, F., Chahal, K.: The state of knowledge distillation for classification. arXiv preprint arXiv:1912.10850 (2019)
25. **Thanh Ma**, Nguyen, T.H., Nguyen, P.K., Thanh-Xuan, N., Thanh-Nghi, D.: ViTIP: AI-Powered Vietnamese Traditional Instrument Preservation System using 3D space. In: International Conference on Intelligent Systems and Data Science (ISDS'24). vol. 2190. Springer (2024)
26. Thúy, N.T.N., et al.: Forms of the spiritual culture living of koho people in lam dong. Journal of Technical Education Science **9**(4), 96–101 (2014)
27. Trier, Ø.D., Reksten, J.H., Løseth, K.: Automated mapping of cultural heritage in norway from airborne lidar data using faster r-cnn. International Journal of Applied Earth Observation and Geoinformation **95**, 102241 (2021)
28. Trochidis, K., Russell, B., Eisenberg, A., Ganguli, K.K., Gomez, O., Plachouras, C., Guedes, C., Danielson, V.: Mapping the sounds of the swahili coast and the arab mashriq: Music research at the intersection of computational analysis and cultural heritage preservation. In: 6th International Conference on Digital Libraries for Musicology (DLfM), Delft, The Netherlands (2019)
29. Van Khê, T.: Vietnamese culture and music. The World of Music **20**(2), 48–52 (1978)
30. Yang, Y., bin Othman, A.N., Hussin, H.B.: Knowledge distillation and transfer learning combined for innovative visualization teaching of non-heritage designs. Applied Mathematics and Nonlinear Sciences (2024)