IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 18 December 2023, accepted 2 January 2024, date of publication 8 January 2024,
date of current version 17 January 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3350880

RESEARCH ARTICLE

Deep Multilevel Cascade Residual Recurrent Framework (MCRR) for Sheet Music Recognition

PING YU[1] AND HAILING CHEN[2]
[1]School of Music, Weifang University, Shandong, Weifang 261061, China
[2]Music and Dance Academy, Heze University, Shandong, Heze 274015, China

Corresponding author: Hailing Chen (talmagerahi@gmx.com)

ABSTRACT Sheet music recognition is a vital technology aimed at converting printed or handwritten musical scores into digital or machine-readable formats. The significance of this technology lies in making music compositions more accessible for editing, performance, learning, and sharing, thereby fostering music education, composition, and culture. It also provides a powerful tool for music analysis, research, and preservation. Our aim is to investigate a sheet music recognition method that offers a simple workflow, high recognition accuracy, and fast model convergence. Specifically, the proposed Deep Multilevel Cascade Residual Recurrent (MCRR) framework for sheet music recognition consists of the following components. Firstly, we introduce additive Gaussian white noise, additive Perlin noise, and elastic deformations such as rotation and stretching to simulate real-world noise in the sheet music images, thereby augmenting the dataset, enhancing model robustness, and mitigating overfitting. Secondly, in the feature extraction phase, we employ a residual Convolutional Neural Network (ConvNet) to address the issue of model degradation and use the multilevel cascade fusion technique to obtain comprehensive feature information, improving the model's feature extraction capability and reducing recognition errors. For note recognition, we use a variant of RNN (Recurrent Neural Network) called SRU (Simple Recurrent Unit), which transforms most computations into parallel processing, speeding up model convergence. Finally, we combine the Connectionist Temporal Classification (CTC) loss function with SRU to eliminate the requirement for strict alignment between data and labels, enabling note classification and recognition. Extensive ablation experiments and comparative analyses, including visual analysis, intuitive illustrations, and quantitative assessments, confirm the effectiveness of the proposed method, demonstrating its superiority over various state-of-the-art methods. The proposed method achieved promising results in both the PrIMus and Camera-PrIMuS datasets. Specifically, in the PrIMus dataset, the method obtained an SeER (Symbol Error Rate) of 1.4571% and a SyER (System Error Rate) of 0.3234%. Notably, it demonstrated high accuracy in pitch, type, and note recognition, scoring approximately 97% in pitch and type accuracy and around 94% in note accuracy. The training time per epoch was relatively low, recorded at 0.56 seconds. In the case of the Camera-PrIMuS dataset, the method achieved slightly lower but still competitive results. It exhibited an SeER of 5.1488% and a SyER of 1.0612%, with pitch and type accuracies around 90%, and note accuracy at approximately 88%. The training time per epoch was slightly higher at 1.93 seconds Furthermore, we compare our method with existing commercial software, namely Capella-scan, PhotoScore, and SmartScore. Among these, Capella-scan delivers the best performance but exhibits lower robustness compared to the proposed method.

INDEX TERMS Sheet music recognition, deep learning, multilevel cascade residual recurrent framework, connectionist temporal classification.

I. INTRODUCTION

Music, as an essential means of cultural expression and communication, relies on detailed records of musical elements such as notes and their related information [1].

The associate editor coordinating the review of this manuscript and approving it for publication was Juan Wang.

One of the primary carriers for this information is sheet music, which provides the most direct way for learning, sharing, and transmitting music [2]. However, many sheets of music have not been made public or published and are still preserved in hardcopy form [3]. These physical copies are susceptible to damage, loss, and deterioration, especially with changing environments and evolving times. Therefore, preserving sheet music in its original physical form holds significant and lasting importance. Traditionally, scanning or photographing sheet music has been a preferred method for preserving hardcopy music sheets. However, this approach is limited by factors such as scan quality and storage capacity. While advances in hardware technology, like scanners and storage devices, have allowed for the storage of clearer, high-quality images of sheet music, computers cannot directly utilize this digitized content. Only when the content of musical symbols in sheet music images is extracted can we flexibly and conveniently harness the power of sheet music, enabling music composition, synthesis, and other operations. The development of computer science and image processing theories and technologies has continuously provided new insights into extracting musical symbols from sheet music images [3]. This has given rise to optical sheet music recognition, a technology that can transform printed sheet music into symbolized music scores, such as MIDI (Musical Instrument Digital Interface) files. This means that optical sheet music images are not just digitized images but files that a computer can ''comprehend.'' This technology has significant and far-reaching implications for music information retrieval, music-assisted teaching, and other domains.

While there are commercial recognition software solutions available, the recognition methods in these tools still face challenges, including poor noise resistance and low recognition accuracy [4]. Thus, there is an urgent need for further research into robust and highly accurate optical sheet music recognition algorithms. Optical sheet music recognition, similar to optical character recognition (OCR), deals with image recognition. However, it faces different challenges in the actual recognition process.

First, most characters have significant differences in shape, making contour recognition the primary method. In contrast, musical notes are nearly identical in shape, consisting mainly of a combination of vertical lines and circles or nearly circular shapes. Recognizing them relies on subtle differences, as shown in Figure 1. For example, compare the grace note, regular note, and accented note. The differences are in the decorations added around the notes, mainly dots, short lines, and greater-than symbols. Even dotted notes and grace notes are distinguished only by the relative position of the dot to the note.

Second, as shown in Figure 2, optical sheet music recognition needs to leverage two-dimensional information extensively. Slight variations in the vertical positions of notes make four different notes represent distinct information because the vertical position of a note determines its pitch. Any deviation can result in incorrect note recognition.

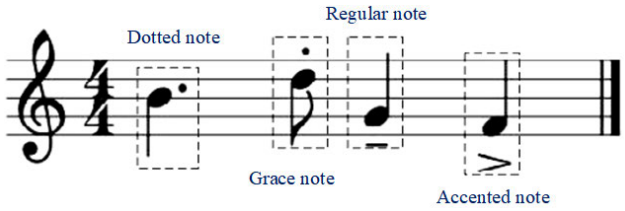

FIGURE 1. Notes at different decorations.

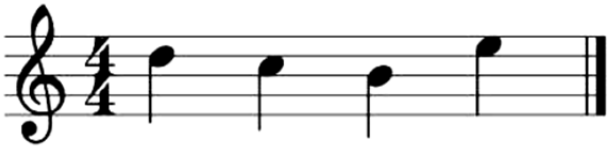

FIGURE 2. Notes at different vertical positions.

Therefore, the difficulty in recognition is no less than OCR. It is evident that sheet music recognition focuses on recognizing musical notes, and this focus differs from OCR. The emphasis is on the feature information carried by the relationships between notes. Therefore, recognizing musical notes, as a distinct category of symbols, holds significant research value.

Since the development of optical sheet music recognition technology, the majority of optimization has been based on traditional frameworks, primarily involving image preprocessing, staff line detection and removal, and symbol recognition and classification. Vo et al. [5] used the Gaussian Mixture Markov Random Field (GMMRF) model for image binarization, effectively removing noise from complex background sheet music while preserving characteristics between notes and staff lines. Rebelo et al. [6] introduced a stable-path spectrum line detection method, lowering the detection error rate to 1.4% without requiring domain knowledge. Wu et al. [7] focused on handwritten sheet music and proposed a method based on multi-dimensional local binary patterns (LBPs) and XGBoost to remove staff lines, achieving improved recognition accuracy with only 0.05% of training data. Calvo-Zaragoza et al. [8] utilized Convolutional Neural Networks (ConvNet) for feature extraction and employed k-Nearest Neighbor (k-NN) and other classifiers for symbol classification, resulting in a reduced classification error rate of 3.61%. While optimizing individual steps within the traditional framework has shown good results, the high complexity and limited overall precision remain challenges. With the continuous development of deep learning, many researchers have adopted end-to-end methods to process entire sheet music images, reducing the complexity of traditional frameworks and avoiding issues with error propagation between individual tasks. Hajic et al. [9] combined ConvNet with bounding box regression for symbol head detection, achieving an accuracy of 81%. However, there are stability issues in classifying early sheet music images with poor binarization and blurry deformation. Choi et al. [10] used a combination of ConvNet and Spatial Transformer Networks (STN) to detect accidentals such as sharps, flats, and naturals

with a detection accuracy of 99.2%. These two methods are limited in scope and have poor scalability as they are tailored to specific symbols. Calvo-Zaragoza et al. [11] combined ConvNet with Bi-directional Long Short-Term Memory (BiLSTM) to form a Convolutional Recurrent Neural Network (i.e., CBLSTM) for recognizing notes throughout the entire sheet music image, resulting in a symbol error rate of 2.16%. However, the model converges slowly, consumes a significant amount of time, and has inadequate recognition accuracy for challenging symbols like grace notes and bar lines. Tuggener et al. [12] connected ResNet-101 with RefineNet upsampling networks and used bounding box detection methods for sheet music image recognition, achieving good results for whole rest symbols but insufficient accuracy (below 50%) for other symbols, particularly accidentals and time signatures.

To address the aforementioned issues, we propose a deep multilevel cascade residual recurrent (MCRR) framework for sheet music recognition. Specifically, the proposed MCRR method begins by augmenting a subset of images in the sheet music dataset using techniques like deformation and noise addition to enhance the model's robustness. In the feature extraction stage, the use of a residual ConvNet effectively mitigates the problem of model degradation. Subsequently, a multilevel cascade pixel-level fusion is applied to features extracted in each layer of the residual ConvNet, consolidating multilayer features into a single feature map, thereby enhancing the model's ability to represent and extract features related to musical notes and consequently improving recognition accuracy. Finally, the feature recognition module employs a Simple Recurrent Unit (SRU) model, which transforms the training process into parallel computations, accelerating model convergence and reducing training time. In the experimental section of this paper, we first validate the effectiveness of the aforementioned optimizations through extensive comparative experiments based on the PriMus (Printed images of music staves) dataset [13] and Camera-PrIMuS dataset [44]. The results demonstrate a significant improvement in symbol error rates, sequence error rates, and noise resistance, with a nearly threefold increase in convergence speed compared to the baseline model. Furthermore, we compare our method with those proposed in existing literature and commercial software, highlighting the advantages in terms of both recognition accuracy and convergence speed.

The innovations and main contributions of this work are as follows:

1) Our framework merges a multilevel cascade residual ConvNet with SRU for optical sheet music recognition. We enhance dataset complexity by augmenting the sheet music image dataset. Utilizing a residual ConvNet during feature extraction mitigates potential gradient vanishing issues. Employing multilevel cascade feature fusion merges feature information from different convolutional layers, amplifying model generalization and feature representation. Finally, SRU, a variant of RNN, expedites model convergence.
2) We present a novel feature extraction network incorporating multilevel cascade feature fusion with residual ConvNet. To capture fine-grained details in challenging musical notes, we integrate information from deep and shallow convolutional layers, enriching feature maps. This fusion method significantly reduces symbol error rates and convergence issues.
3) Our proposed note recognition network combines SRU with Connectionist Temporal Classification (CTC) functions. RNN and CTC alignment effectively construct notes, with SRU resolving long-range dependency issues and accelerating convergence. Experimentally, SRU achieves nearly three times the convergence speed of LSTM networks while maintaining accuracy.
4) We rigorously validate our framework on the test set through ablation experiments, affirming the effectiveness of image preprocessing, residual ConvNet, multilevel cascade fusion, and SRU modules. Comparative analysis with existing methods and commercial software demonstrates our framework's superior accuracy and robustness.

Our arrangement for the subsequent sections of the article is as follows:

Section II covers the related work. It elaborates on the background and significance of sheet music recognition, provides a detailed analysis and comparison of the two major research approaches for this task, and the outstanding issues. Section III focuses on the improvement of our sheet music recognition method. This section provides detailed explanations of various aspects, including image preprocessing, the structure of the residual ConvNet, multilevel cascade feature fusion, SRU, and the CTC function. Section IV analyzes and validates the effectiveness of the model through extensive experiments. This includes assessing the impact of image enhancement techniques, the improvements brought by the residual ConvNet, the supplementary information provided by multilevel cascade feature fusion, and the optimization effect of the SRU. Furthermore, the paper compares the recognition performance of our method with existing methods and commercial software to highlight the advantages. Sections V and VI provide a concise summary of our work and offer insights into future research directions for sheet music recognition methods.

II. RELATED WORKS

Research in the field of optical sheet music recognition was initiated by Fujinaga et al. [14] in 1988. Although it has gone through many developmental stages, progress has been slow, and some technologies are still not mature enough. In 1992, Blostein et al. [15] provided a summary of the technologies existing in the optical sheet music recognition system by reviewing its development over nearly thirty years. In 2001, Bainbridge and Bell [16] proposed a generic framework-based optical sheet music recognition algorithm that decomposed the technology into four subtasks: image preprocessing, note recognition, music information

reconstruction, and final representation construction. This led to some breakthroughs and laid the foundation for subsequent research. Following this, Homenda [17] and Rebelo et al. [18] introduced pattern recognition methods applicable to musical symbols, while Jones et al. [19] presented a research approach for musical score images, covering aspects like digitization, recognition, and restoration. They also summarized the software and hardware required for optical sheet music recognition systems and explored potential applications, advancing the research of optical sheet music recognition. In recent years, the extensive use of big data has driven researchers toward data-driven recognition methods for completing various subtasks. Due to the outstanding performance of deep learning in image processing, researchers have started using neural network models for optical sheet music recognition. They are gradually adopting end-to-end optical sheet music recognition methods, simplifying the generic framework-based research and providing new perspectives and angles for optical sheet music recognition research.

A. UNIVERSAL FRAMEWORK-BASED METHODS

Early universal frameworks divided the optical sheet music recognition process into four sub-tasks: image preprocessing, note recognition, note reconstruction, and symbol representation of the music, as shown in Figure 3. The first step is image preprocessing, which separates useful information, such as notes and staff lines, from background noise to reduce its impact on the recognition. The next step aims to isolate staff lines and notes by removing the staff lines from the music score, preserving the relative position information between staff lines and notes. The final step involves note reconstruction and the creation of the final representation. This phase mainly deals with the reconstruction of recognized information to determine the correctness of note sequences and semantic information, ensuring alignment with musical conventions.

1) IMAGE PREPROCESSING

Image binarization is a critical step in most image analysis and processing systems. In cases where the image quality is poor, noise reduction, enhancement, and offset correction methods are employed during image preprocessing to make subsequent recognition more efficient and robust. Almost all optical sheet music recognition systems start with image binarization, converting grayscale images into binary images. Traditional binarization methods screen pixels based on the characteristics of information in the image. These methods can be broadly categorized as global thresholding methods and adaptive thresholding methods. Global thresholding applies a fixed threshold to the entire image and is preferred for its simplicity and computational efficiency. However, it often performs poorly on non-uniform backgrounds. In contrast, adaptive thresholding methods, like those based on the Otsu algorithm, adaptively choose thresholds based on local image characteristics and are considered faster and more effective. Background complexity in sheet music can limit the effectiveness of traditional binarization methods. Methods

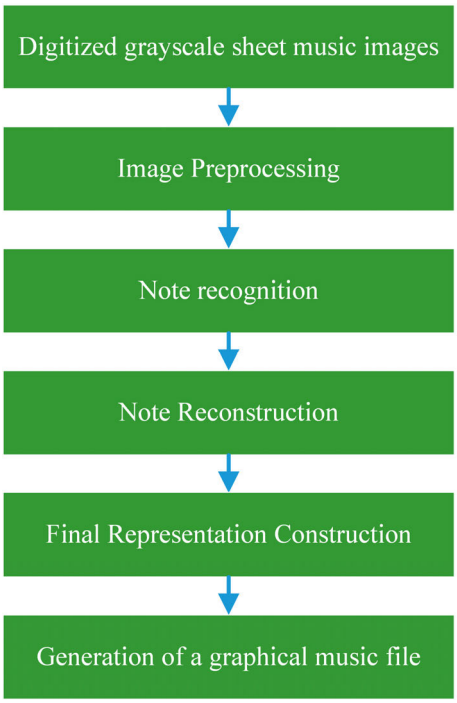

FIGURE 3. Universal framework schematic.

like [20] have integrated the contents of the sheet music into a GMMRF (Gaussian Mixture Markov Random Field) model, creating an automatic binarization system. This approach effectively removes noise from complex music scores while preserving the characteristics of notes and staff lines.

2) NOTE RECOGNITION

Note recognition follows a hierarchical approach. First, it separates the background from the sheet music by detecting and removing the staff lines, which isolates the staves from the notes. Next, it segments the notes based on their characteristics, extracts various types of note elements, and performs recognition and classification.

Although staff lines provide auxiliary information about the pitch and other details of the notes, their presence can interfere with note recognition. Therefore, the accuracy of staff line detection and removal directly impacts note recognition accuracy. The overlapping nature of staff lines and notes makes staff removal challenging. When staff lines are not removed accurately, the remaining lines can interfere with note recognition. Conversely, if staff removal is excessive, it may lead to fragmented note shapes, which also pose challenges for note recognition.

Staff line detection is typically divided into methods based on statistical transformations and methods based on structural search. Methods based on statistical transformations use techniques such as horizontal projection, Hough transform, and wavelet transform for staff line detection. While these methods offer strong noise resistance, they are susceptible to deformation in music sheet images. Structural search-based methods prioritize deformation resistance over noise

resistance. For instance, Hatori et al. [21] utilize dynamic programming with feature points. Rebelo and Cardoso [6] propose a stable path staff line detection method, reducing detection errors to 1.4% without the need for domain knowledge. Balancing noise and deformation resistance remains a challenge, as music sheet image quality can be challenging to control.

In staff line removal, techniques like skeletonization are used to eliminate staff lines. Bui et al. [22] employ Line Adjacency Graph (LAG). Martin [23] use vector lines. Konwer et al. [24] present a method based on multidimensional LBPs and XGBoost to remove staff lines, achieving model accuracy with only 0.05% of training data. However, Pugin et al. [25] omit staff line removal and directly apply Hidden Markov Models (HMM) for note recognition, increasing computational requirements and data modeling difficulty while achieving subpar results.

In a general framework, note recognition involves the segmentation of notes into elementary components. Common note elements include noteheads, stems, tails, and beams, which can have various shapes, including straight lines, ellipses, and curves. Segmentation methods need to be adapted based on the individual characteristics of note elements. For beam segmentation, region growing and improved Hough transforms are used. Other notes are re-extracted using template matching. Note segmentation and classification are often combined into a single step, although some researchers separate these processes. Reed et al. [26] divide segmentation into three stages: first, they detect lines and curves using the LAG method; second, they use a character contour method to detect accidentals, rests, and clefs; finally, they detect noteheads using template matching. Mahoney [27] construct a set of candidate options for one or multiple classes of symbols and select matching candidates based on relevant descriptors. Homenda and Sitarek [28] study decision tree and clustering methods for note feature extraction. Pacha et al. [29] propose a rule-based system for note representation controlled entirely by grammar. Pugin et al. [30] propose a segmentation method based on HMM that simultaneously handles note recognition. However, this method is applied to relatively simple sheet music without tied notes. Taubman et al. [31] utilized statistical moments for note recognition. Rebelo et al. [32] conduct a comparative evaluation of four recognition and classification methods. Among these methods, Support Vector Machine (SVM) and k-NN algorithms demonstrated better performance, while HMM and Neural Networks (NNs) did not exhibit remarkable classification results. The effectiveness of these four methods varies significantly depending on the types of note elements.

3) NOTE RECONSTRUCTION AND FINAL REPRESENTATION CONSTRUCTION

Due to the strong interrelationships between notes, note reconstruction requires the integration of contextual information. This goes beyond staff information and encompasses other elements in the music sheet, such as time signatures, key signatures, clefs, and more. Grammar-based reconstruction effectively addresses this issue. Bainbridge and Bell [33] use Definite Clause Grammars (DCGs) to specify relationships between recognized music shapes, outlining the CANTOR system. This system selects the most likely notes based on the symbol pack contents. Bainbridge and Bell [34] enhance the CANTOR system by incorporating feature positions *(x, y)*. Rossant and Bloch [35] introduce an image rule-based optical sheet music recognition system using graphical rules to ensure note consistency, such as fixed distances between accidentals and noteheads.

B. DEEP LEARNING-BASED METHODS

With the rapid development of computer vision and the significant achievements of neural networks in image processing, researchers have begun to apply deep learning to optical sheet music recognition. This is achieved by constructing specific models and training them on a large dataset of sheet music images to learn model parameters for specific tasks.

Some research work has applied neural networks to various subtasks within the universal framework, including tasks like staff removal, note recognition, and note classification, utilizing ConvNets. Calvo-Zaragoza et al. [8] treated staff detection as a classification task and employed ConvNets for staff detection. They labeled each pixel as staff or note and trained the model using paired data with and without staves. Experimental results showed that even without post-processing, this approach outperformed many traditional methods. Pinheiro et al. [36] used ConvNets for note recognition and compared it with classic networks like LeNet, AlexNet, and GoogleNet, with GoogleNet showing the best performance in note recognition. Rebelo et al. [18] were the first to introduce neural networks in the note classification stage, laying the foundation for further research. Subsequent work by Zhou and Jia [37] involved using multilayer perceptrons to build targeted models for two types of notes, improving classification results. In a study mentioned in [38], ConvNet models were employed for staff removal and note classification. Among these, the ResNet model worked best for undistorted images with 25 convolutional layers and nearly 5 million parameters. For images with inconsistent sizes, the VGG model was used, offering similar recognition rates with only 13 convolutional layers but including 8 million parameters.

Some research takes an end-to-end approach, offering new perspectives for the complex and diverse subtasks. This approach involves training on entire music sheets as inputs, rather than learning individual symbols, simplifying the research based on the universal framework. For instance, Hajič et al. [9] combined ConvNet with bounding box regression for notehead detection, achieving an accuracy of 81%. However, it had issues with classification stability when dealing with early low-quality music sheet images suffering from poor binarization and blurry deformation. Choi et al. [10] employed ConvNet and STN to detect

TABLE 1. The summary of deep learning-based methods.

Study	Methods	Dataset	Advantages	Disadvantages
Calvo-Zaragoza et al. [8]	Staff detection using ConvNets	ICDAR 2013 Competition on Music Scores	Automated staff-line removal, enhanced readability of music scores, ConvNet's pattern recognition capabilities	Computational resource-intensive, effectiveness reliant on dataset diversity and quality, handling variations in notation styles & complex layouts
Pinheiro et al. [36]	Deep learning for handwritten musical symbols recognition	HOMUS	Improved recognition accuracy, ability to interpret handwritten symbols	Computational complexity, dataset quality impacts performance
Zhou et al. [37]	Optical sheet music recognition using combined neural networks	A data set of both real handwritten scores and scanned scores	Integration of neural networks for optical sheet music recognition, potentially higher accuracy in recognizing music elements	Potential complexity in model architecture, dataset dependency
Pacha et al. [38]	Self-learning optical sheet music recognition	MUSCIMA & Pascal VOC Challenge 2006	Adaptability and self-learning capabilities, potential for improving optical sheet music recognition performance through continuous learning	Lack of details regarding specific methods and datasets
Hajič et al. [9]	ConvNets for detecting noteheads in handwritten scores	MUSCIMA++	Utilization of ConvNets for accurate notehead detection, bounding box regression for localization	Model complexity, dependency on diverse and high-quality training data
Choi et al. [10]	Bootstrapping samples for ConvNet-based accidental detection	imslp.org	Improved accidental detection using ConvNet, leveraging bootstrapping for dataset enhancement	Relies on specific piano score datasets, potential bias due to dataset selection
Sober-Mira et al. [39]	Pen-based music document transcription	self-collection	Utilization of pen-based music data, potential for accurate transcription of handwritten music	-
Calvo-Zaragoza et al. [11]	End-to-end optical sheet music recognition with Neural Networks	52 Common Western Music Notation symbols	End-to-end optical sheet music recognition, potential for comprehensive music recognition using neural networks	Computational intensity, reliance on dataset quality and diversity
Tuggener et al. [17]	Deep watershed detector for music object recognition	MUSCIMA++	Deep watershed approach for accurate music object detection, potentially improved recognition accuracy	Potential complexity in implementation, performance dependency on dataset quality

TABLE 1. *(Continued.)* **The summary of deep learning-based methods.**

Martinez-Sevilla J C et al. [45]	Holistic approach for aligned music and lyrics transcription	Aligned Music Notation and Lyrics Transcription (AMNLT)	Integration of music and lyrics transcription, potential for precise alignment and transcription of music and lyrics	Computational complexity, potential challenges in alignment with diverse datasets
Castellanos F J et al. [46]	Neural approach for full-page optical sheet music recognition of Mensural documents	CAPITAN & SEILS	Neural network-based full-page optical sheet music recognition, potential for accurate recognition of Mensural notation	Reliance on specific Mensural music dataset, computational demands for model training
Ríos-Vila A et al. [47]	End-to-end full-page optical sheet music recognition for Mensural Notation	Il Lauro Secco & CAPITAN	Comprehensive end-to-end optical sheet music recognition for Mensural notation, potential for accurate recognition	Model complexity, dataset dependency
Ríos-Vila A et al. [48]	Complete optical sheet music recognition via agnostic transcription and machine translation	Il Lauro Secco & CAPITAN	Comprehensive optical sheet music recognition approach via transcription and translation, potential for versatile recognition of various music notations	Potential complexities in translation, reliance on diverse and high-quality data for effective transcription and translation

accidentals, sharps, flats, and naturals, achieving 99.2% accuracy. These methods focus on recognizing specific symbols and have limited applicability and scalability. In a study mentioned in [39], electronic pen technology and ConvNet models were used to convert handwritten music sheets into electronic formats, but it also highlighted the importance of related feature characteristics. Calvo-Zaragoza et al. [11] combined ConvNet and BiLSTM to create a CBLSTM to recognize notes throughout an entire music sheet image, achieving a symbol error rate of 2.16%. However, this model had slow convergence, high time consumption, and less accuracy in recognizing challenging symbols like grace notes, accidentals, rests, etc. The outstanding performance of R-CNN, YOLO, and SSD in object detection led Tuggener et al. [17] to consider region-based object detection for optical sheet music recognition, connecting ResNet-101 with RefineNet upsampling networks and combining them with bounding box detection methods to recognize music sheet images. While it performed well in recognizing whole rests, its accuracy in recognizing other notes, especially accidentals and time signatures, fell below 50%. Martinez-Sevilla et al. [45] introduced the Aligned Music Notation and Lyrics Transcription (AMNLT) challenge, aiming to extract content from vocal music document images. Despite existing methods dealing with music notation and text, they lacked proper alignment, crucial for retrieving vocal music content. Their proposed holistic neural approach showedover 80% relative improvements in transcription and alignment evaluation metrics. Castellanos et al. [46] proposed a full-page optical sheet music recognition system for Mensural notation scores, leveraging selectional auto-encoders and convolutional recurrent neural networks. Their method extracted symbolic music information effectively, demonstrating successful behavior on two Mensural collections. Ríos-Vila et al. [47] presented a segmentation-free full-page optical sheet music recognition system for transcribing page images into music notation, alleviating manual labeling efforts. The methodology showed promising results, especially beneficial for early music written in mensural notation. In addition, They focused on the integration of recognition results into practical standard music formats for end-users' benefit [48]. They proposed adding machine translation systems to the recognition pipeline, providing a feasible solution for complete optical sheet music processes, especially in scenarios with limited training data. The main methods used for these methods, the data sets used, and their advantages and disadvantages are summarized in Table 1.

C. OUTSTANDING ISSUES

While image processing techniques have seen rapid advancements, optical sheet music recognition has been progressing

slowly despite years of effort. As of now, current algorithm research continues to face the following challenges:

1) The first challenge lies in the complexity of algorithms within the universal framework. Each step involves its own set of difficulties. Staff detection requires a balance between noise resistance and deformation resistance, considering both statistical transformation and structural search-based methods. Staff removal increases the difficulty of recognizing dotted notes. Note recognition and classification require different methods depending on the specific characteristics of the notes. Finding a universal algorithm is challenging, and classification performance varies significantly for different types of notes. These issues result in an overall lack of accuracy in optical sheet music recognition tasks.

2) The second challenge is related to the lack of richness and diversity in note sequences within the dataset. This deficiency hampers the model's generalization capabilities and often leads to overfitting. A clean dataset can negatively affect algorithm robustness, resulting in decreased recognition accuracy when dealing with noisy or deformed music sheet images.

3) The third challenge is introduced by deep neural network algorithms trained end-to-end. While they simplify the complexity of the universal framework, they no longer allow separate analysis and research of key steps. This reduction in the number of steps could decrease error propagation, but optical sheet music recognition tasks are highly sensitive to fine details, particularly in the recognition of challenging notes. Inadequate feature extraction capabilities significantly impact recognition accuracy.

4) The fourth challenge pertains to the slow convergence rate of the note recognition method using BiLSTM models. Increasing the number of parameters extends the training time, making it time-consuming to retrain the model with each parameter modification.

To address these challenges, this paper proposes an optical sheet music recognition framework that aims to streamline the entire process, enhance recognition accuracy, optimize convergence speed, and reduce training time.

III. THE PROPOSED METHOD

As previously mentioned, the mainstream algorithms for optical sheet music recognition can be categorized into two approaches: those based on a universal framework and those based on deep learning. Universal framework-based optical sheet music recognition algorithms divide the entire task into individual subtasks, each of which is complex. While the recognition accuracy in each step can be improved through algorithm optimization, the error propagation nature directly affects the subsequent recognition of staves and lacks a unified approach for note recognition, requiring tailored recognition methods based on note characteristics. On the other hand, deep learning-based algorithms effectively avoid such issues by processing the entire image as input. A universally applicable method can be obtained through end-to-end parameter learning and model training. However, these approaches face challenges in terms of recognition accuracy and slow model convergence, leading to longer training times. Therefore, this paper introduces a deep learning-based algorithm with strong feature recognition capabilities and faster model convergence.

A. OVERVIEW

We have proposed the deep multilevel cascade residual recurrent framework for sheet music recognition based on a deep learning architecture. The overall process, as shown in Figure 4, consists of three main components: image preprocessing, feature extraction, and note recognition.

First, the input sheet music image is resized to a fixed height of 128 pixels, with its width scaled proportionally. We introduce additive Gaussian white noise, additive Perlin noise, and elastic deformation to simulate various less-than-ideal sheet music images encountered in real-world scenarios.

Subsequently, we utilize a five-layer deep residual ConvNet to extract features related to musical notes within the image. Features extracted at different levels are fused into a single feature map in a cascade manner, where high-level semantic information is integrated with low-level details through pixel-level fusion. This cross-level information exchange results in more comprehensive features, enhancing the quality of information available for the subsequent note recognition stage.

Finally, the extracted feature sequence undergoes dimension transformation to serve as input for the note recognition segment. Note sequences are recognized using BiSRU (Bi-directional Simple Recurrent Network), and note classification is achieved through the use of CTC functions, which do not require forced alignment of data in the dataset.

1) IMAGE PREPROCESSING

In deep learning applications, the quantity and quality of the training dataset directly impact the learning of model parameters. A small dataset can lead to insufficient learning of some features, making the model prone to overfitting, resulting in a significant drop in accuracy on the test set. On the other hand, a fixed training dataset quality can limit the model's ability to recognize deviations from quality in real-world data, resulting in lower recognition accuracy. To ensure that the proposed model performs well on low-quality inputs and various types of sheet music, this paper introduces computer-based techniques to simulate real-world noise, such as additive Gaussian white noise, additive Perlin noise, and elastic transformations, to create various types of less-than-ideal sheet music images, thereby expanding the dataset, enhancing the model's robustness, and avoiding overfitting issues.

Additive Gaussian white noise is a commonly used data augmentation technique, which introduces random deviations with a normal distribution to pixel values, following a uniform distribution on the power spectral density. The noise mean μ is the same as the pixel mean for the entire dataset, while the standard deviation for the training set is chosen from the standard deviations of the pixel values in the original dataset. Take the music score in Figure 5(a) for example,

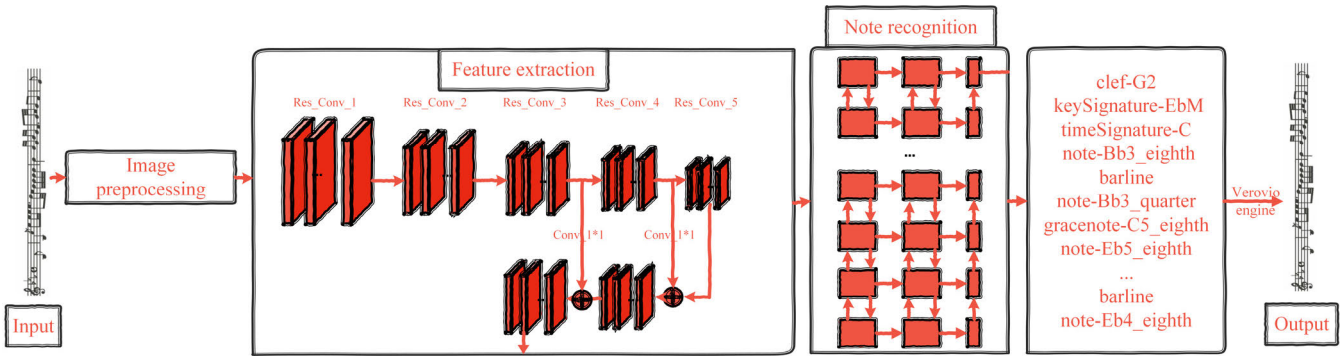

FIGURE 4. The proposed MCRR framework.

adding Gaussian white noise simulates low-quality printing or scanning of sheet music images, as seen in the result in Figure 5(b).

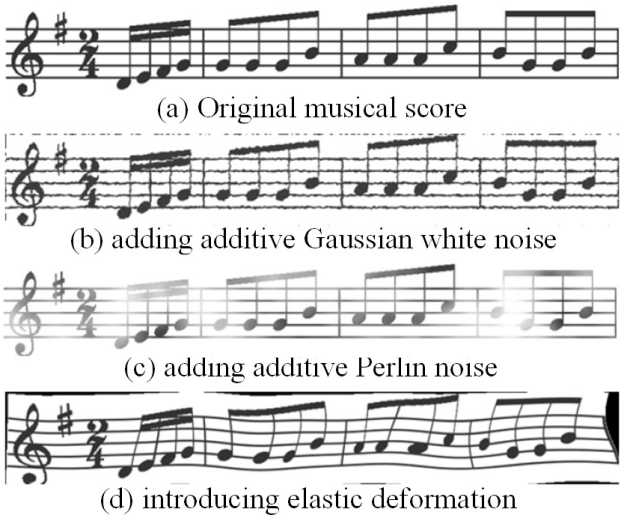

FIGURE 5. Visual results of three data processing methods for simulating less-than-ideal sheet music images.

Additive Perlin noise generates gradient noise. Compared to additive Gaussian white noise, it produces a larger range of lighter or darker noise in the image. Figure 5(c) displays the sheet music with Perlin noise added to the original sample. This results in localized fading, even causing some notes to appear faded, making the staff lines less apparent. It can also generate large areas of dark regions, simulating cloud-like noise that occurs in sheet music images due to uneven ink distribution or long-term storage. The average size of the lighter or darker noise is determined by the frequency parameter, which is a random value between the pixel size of the notes and the average width of a complete measure.

Elastic deformation applies a smooth local random affine transformation field to generate wave-like displacements. It applies numerous affine and geometric transformations, such as rotation, skewing, compression, and stretching to each image, enhancing the diversity of the data without the need for manually defining geometric transformations. In Figure 5(d), elastic deformation is applied to the image, simulating minor folding and distortion that can occur during the printing process. The strength factor σ controls the degree of distortion, with larger values resulting in less distortion, while the smoothness factor α controls the degree of deformation, with larger values leading to deformation closer to linear transformation. By using these three data augmentation methods, the dataset contains high-quality sheet music images as well as noisy and distorted sheet music images.

2) RESIDUAL ConvNet STRUCTURE FOR NOTE FEATURE EXTRACTION

The sheet music images exhibit discrete and fairly even distribution of notes. They primarily consist of solid or hollow circular shapes, with linear or curved structures from multiple directions. Some notes share the same shape, differing only in their positions. Additionally, certain notes are relatively small, making them susceptible to confusion with noise in the sheet music. To address these characteristics, ConvNet is employed for extracting features from the sheet music images. The ConvNet's convolutional layers have local connections and weight sharing properties, which facilitate the extraction of note edge features and positional information. The activation function layer enhances the ConvNet's expressive power, allowing it to be differentiable and achieve a nonlinear mapping from low-dimensional simple features to high-dimensional complex features within the sheet music images. The pooling layer reduces the number of weight parameters, accelerates computation, and prevents overfitting while retaining the primary features extracted by the convolutional layers. Typically, to enhance model accuracy, one might increase the width or depth of the ConvNet. However, this can lead to issues like gradient vanishing/exploding during parameter updates, resulting in non-convergence. Therefore, we used Residual convnet structure for note feature extraction [40], as illustrated in Figure 6.

3) DEEP MULTILEVEL CASCADE STRUCTURE FOR NOTE FEATURE EXTRACTION

In the process of feature extraction using ConvNet, the number of convolutional layers increases to extract features at different levels. Generally, shallow features for notes include information about their positions and edges. Although deep

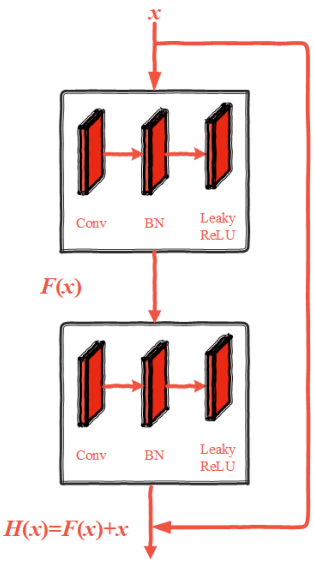

FIGURE 6. Our residual structure.

features have smaller resolutions, they contain rich semantic information, which can aid in better note recognition. However, notes are typically composed of fixed geometric shapes with minimal variations between them. Even notes with the same shape may differ in pitch, representing distinct notes. Therefore, note recognition relies on enhancing the recognition accuracy through differences in details. If only the deepest ConvNet layer output is used for note recognition, it may lack the details obtained from the shallower network layers, which could impact the recognition accuracy. Hence, in the feature extraction process for notes, a multilevel cascade fusion of deep semantic information and shallow detail information is performed using the ConvNet. This approach enriches the feature set for note recognition, allowing for more comprehensive and detailed information in the subsequent note recognition process, as illustrated in Figure 7.

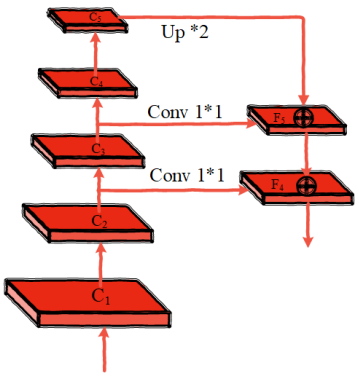

FIGURE 7. Our deep multilevel cascade structure.

In Figure 7, the left part features five layers of residual-style ConvNets, which extract features from the sheet music images from bottom to top. Each residual-style ConvNet layer internally undergoes two convolutions, with max-pooling (MP) following each convolution. The original sheet music image is sequentially processed through each residual-style ConvNet layer, resulting in feature maps C_1, C_2, C_3, C_4, and C_5. Their sizes gradually reduce to 1/4 of the size of the previous layer. The convolution kernel size for each layer is 3*3, with the number of convolution kernels being 32, 64, 128, 256, and 256, as indicated in Table 2.

The right part of the structure is the top-down feature fusion section. It performs pixel-level cascade fusion between feature maps containing semantic information from the deeper layer C_5 and the previous layer feature map C_4. To ensure successful fusion, C5 is upsampled by a factor of 2, making its size consistent with C_4. C_4 undergoes convolution with a 1*1 kernel, ensuring its dimension matches the upsampled C_5. The fused feature map, F_5, is obtained. The same operation is applied to feature maps F_5 and C_3, resulting in feature map F_4. This process realizes multilevel feature fusion, ensuring that the feature vector used in the subsequent note recognition contains more comprehensive and detailed information.

4) SIMPLE RECURRENT UNITS (SRU) FOR NOTE RECOGNITION

While several neural network models can effectively recognize notes in sheet music images, the sequential nature of note sequences, where the types of notes and their order are fixed for each piece of sheet music, presents a challenge. The relationship between notes at the current time step and those in the preceding and subsequent time steps is strong, and any changes in this relationship indicate a change in the musical information, possibly leading to recognition errors. Given the robust recognition capabilities of RNNs for sequential data, RNNs are employed for note recognition. During training, RNNs often face the issue of gradient vanishing due to the large data lengths. Most modern RNN structures use the gate mechanisms, such as LSTM or GRU, to control information flow and mitigate potential gradient vanishing issues. However, LSTM & GRU models, including their forget gates, input gates, and cell states, still depend on the output of the previous time step's hidden unit, limiting parallel processing speed to a significant extent. Therefore, We use the SRU module, as shown in Figure 8, to overcome the constraints imposed by the compulsory continuity between time steps.

The differences between LSTM and SRU are illustrated in Figures 9. In Figure 9(b), calculations within the box can occur simultaneously, substantially reducing the computation load compared to Figure 9(a). Moreover, for input sequences of equal length at a specific time step and SRU and LSTM models of the same dimensions, LSTM's weight dimensions increase as the input length grows, significantly increasing computational load during hyper-parameter training. In contrast, SRU's weight dimensions remain relatively small, resulting in reduced computational load during training and accelerating model computation speed.

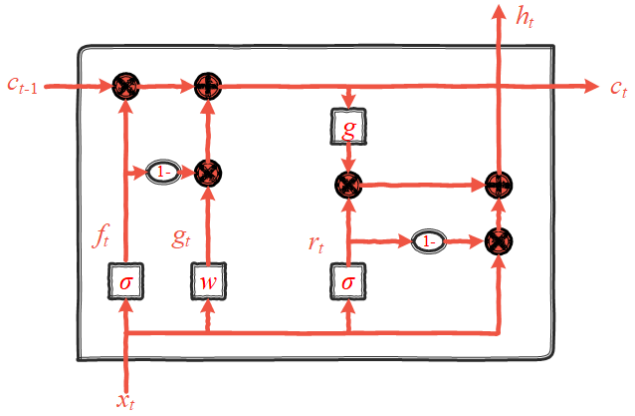

FIGURE 8. SRU detailed structure.

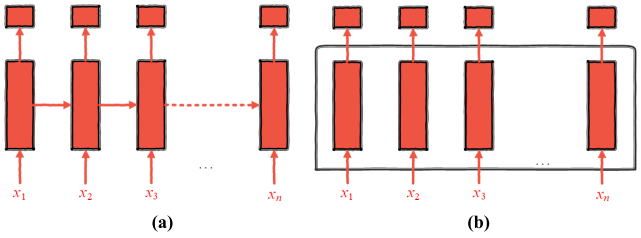

FIGURE 9. The differences between LSTM and SRU.(a) LSTM structure. (b) SRU structure.

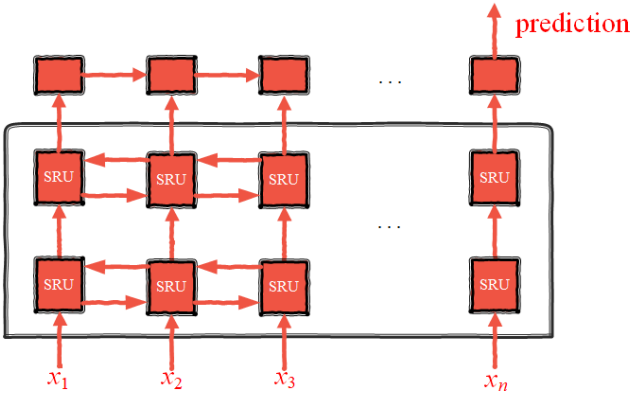

FIGURE 10. BiSRU structure.

Therefore, the process of note recognition is illustrated in Figure 10. The model comprises two layers of BiSRU. Due to the fixed height of 128 pixels for each sheet music image in the image preprocessing stage as shown in Table 2, and the constant number of convolutional kernels selected in the feature extraction network, the number of hidden units is fixed at 512 in each SRU. In each SRU, both forward and backward propagation weight calculations are carried out by 512 hidden layer units. In Figure 10, each rectangle within the box represents an SRU that performs matrix computations for the current time step input, while each rectangle outside the box represent the dot product calculation units between the outputs of previous and subsequent time steps, using this structure to facilitate most parallel computations and expedite the model's convergence speed.

TABLE 2. Parameter configuration for each layer.

Module	Name	Size
Input	Input	128*weight*1
Feature Extraction	C_1	3*3*32
	MP_1	2*2*32
	C_2	3*3*64
	MP_2	2*2*64
	C_3	3*3*128
	MP_3	2*2*128
	C_4	3*3*256
	MP_4	2*2*256
	C_5	3*3*256
	MP_5	2*2*256
Note recognition	BiSRU	512
	BiSRU	512
	CTC	1780

5) CONNECTIONIST TEMPORAL CLASSIFICATION (CTC) FOR NOTE RECOGNITION

When training RNN to model the temporal aspects of sheet music image data, the network must provide an expected output for each note in the sequence, which corresponds to specific labels. However, in the loss calculation process, RNN requires a strict alignment between note-to-label correspondence and the original image pixels. Without proper alignment, pre-segmentation of input data or post-processing of output data is needed. Both manual alignment and the use of open-source tools are time-consuming and prone to alignment errors, significantly impacting recognition accuracy, especially in the case of sequence data. Therefore, We employ CTC loss function, as proposed by Graves et al. [41], as a replacement for the cross-entropy loss function. The CTC loss function is the optimal choice when working with RNN models for sequential data, and experimental results have shown that networks composed of CTC in combination with BiLSTM outperform networks constructed with RNN and HMM [42]. Unlike other loss functions, the CTC loss function allows model training and parameter learning on unaligned datasets, focusing solely on the relative accuracy of label positions. This eliminates the need for forced alignment, significantly reducing the requirements on the training dataset. Specifically, CTC transforms the network's output into a conditional probability distribution over label sequences. When the conditional probability distribution is determined, the network selects the most likely labels for a given input sequence, maximizing the probability of the label sequence, thus achieving the final target sequence. In the case of a given sheet music image, the probability of outputting notes varies. Each position in the note sequence constitutes a selectable output path, and the different conditional probabilities for output notes result in varying path probabilities. Selecting the note sequence with the highest probability increases the likelihood of outputting the correct sequence. By traversing multiple

paths and selecting the one with the highest probability, precise note recognition and classification are achieved.

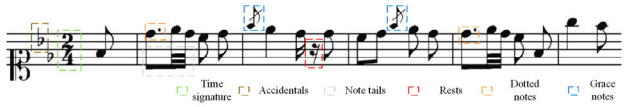

FIGURE 11. Examples of notes in the PrIMus dataset.

FIGURE 12. Examples in the Camera-PrIMus dataset.

IV. EXPERIMENT

A. DATASET

The first experimental data were sourced from the open source PrIMus dataset [13], comprising approximately 87,687 real music scores. As shown in Figure 11, each example consists of single-staff five-line music notation, divided into approximately 4 to 7 measures. Most examples contain not only simple arrangements of musical notes but also a variety of musical symbols, including clefs, time signatures, accidentals, rests, and challenging symbols such as grace notes and dotted notes. Unless otherwise specified, the default baseline dataset is the PrIMus dataset. In addition, for the robustness evaluation of proposed method, we utilized the Camera-based Printed Images of Music Staves (Camera-PrIMuS) database [44]. This dataset consists of 87,678 genuine music staves of monophonic incipits extracted from the Repertoire International des Sources Musicales (RISM). Unlike the PrIMuS dataset, which primarily contains scanned images of printed music scores, the Camera-PrIMuS dataset incorporates real incipits captured using cameras, portraying various realistic scenarios and potential distortions in the music sheets. Each incipit within the Camera-PrIMuS database offers diverse representations: an image depicting the score (both in plain form and with artificial distortions, as shown in Figure 12), multiple encoding formats for symbol information, and a MIDI file representing the musical content. This study augmented some examples with additional Gaussian white noise, additive Bernoulli noise, stretching, rotation, and elastic deformations to enhance the two datasets. The datasets were then divided into training, validation, and test sets in an 8:1:1 ratio.

B. EVALUATION METRICS

Currently, there is no unified standard for evaluating optical sheet music recognition algorithms, and different researchers use their own evaluation metrics. We employ a range of metrics, including Sequence Error Rate (SeER), Symbol Error Rate (SyER), pitch accuracy, type accuracy, and note accuracy, for comprehensive evaluation.

SeER: It is the ratio of the number of incorrectly predicted sequences to the total number of sequences. A sequence is considered incorrect if it contains at least one error in terms of notes, pitches, rests, and more.

SyER: It measures the ratio of the average number of insertions, modifications, deletions, or other editing operations required to produce the label sequence from the predicted sequence concerning the current sequence length.

Pitch Accuracy: the proportion of notes whose pitch is correctly predicted to the total number of notes.

Type Accuracy: the proportion of notes whose type is correctly predicted to the total number of notes.

Note Accuracy: the proportion of notes whose pitch as well as type is correctly predicted to the total number of notes.

There is no absolute correlation between SeER and SyER, while the former describes the error proportion of centralized test examples, the latter summarizes errors in the musical symbols within the examples. This study focuses more on evaluating symbol accuracy as a measure of note recognition precision, but SeER still holds significance in various applications.

By contrast, pitch accuracy, type accuracy, and note accuracy metrics offer specialized assessments in optical sheet music recognition task, focusing on distinct facets of musical transcription. Unlike SeER and SyER, these metrics delve deeper, evaluating the precision in identifying pitch, distinguishing between various musical symbol types, and accurately transcribing complete musical notes with their attributes. While SeER and SyER provide broad error rates, the finer granularity of pitch, type, and note accuracy metrics allows for a more nuanced evaluation, catering to specific components crucial in achieving accurate music recognition.

C. EXPERIMENT SETTINGS

The experimental environment for this study was as follows: Ubuntu 18.04 operating system, Intel Core i7-10700 CPU, 32GB RAM, Nvidia GTX 2080 GPU, and the Pytorch deep learning framework. The model used the Adam adaptive learning rate algorithm for optimization, combining momentum-based and adaptive algorithms, with an initial learning rate set to 0.001. The batch size was set to 32. After every 1000 iterations, the algorithm assessed the symbol error rate on the validation set to verify the model's accuracy. The entire model underwent approximately 64,000 iterations.

Furthermore, we utilized the Verovio tool for symbolic representation and rendering of music scores, capable of presenting and rendering musical scores in symbolic form as visualized score images. This tool is commonly used to convert symbolic information of musical scores into computer-visualized images, rather than dealing with graphic representations. The general steps for using the Verovio tool to generate the final visual representation involve initially parsing the symbols obtained from music score recognition,

including notes, pitches, time signatures, key signatures, etc., recognized using proposed recognition method, into a computer-understandable MEI (Music Encoding Initiative) data format. Then, leveraging the Verovio tool, we rendered the musical data into SVG (Scalable Vector Graphics) format. Finally, we used the svg2png function within the Pillow library to obtain PNG (Portable Network Graphics) format score images suitable for visual display.

D. ABLATION EXPERIMENTS

1) EFFECTIVENESS OF DATA AUGMENTATION

As the proposed method is an improved version of the CBLSTM model [11], we first assessed the impact of the augmented dataset on the experimental results of the CBLSTM baseline model. The dataset was augmented using methods such as additive Gaussian white noise, additive Perlin noise, and elastic transformations like rotation and stretching. The model is trained on augmented datasets and original datasets, and their SeER and SyER are compared using the same test set. Table 3 illustrates the results concerning the PrIMuS dataset, while Table 4 delineates the outcomes pertaining to the Camera-PrIMuS dataset.

TABLE 3. The comparison performance about ablation baseline models for data augmentation in the primus dataset.

Data Augmentation	SeER/%	SyER/%
Before	19.9023	2.9634
After	14.3498	3.2480

TABLE 4. The comparison performance about ablation baseline models for data augmentation in the camera-primus dataset.

Data Augmentation	SeER/%	SyER/%
Before	37.0387	10.3500
After	25.6667	11.9034

From Tables 3 & 4, it is evident that models trained on the augmented dataset exhibit a noticeable decrease in the SeER metric. This experimentation demonstrates that data augmentation effectively helps recognize images under different lighting conditions and various printing qualities. However, there was no improvement in the SyER metric, and in some cases, recognition accuracy decreased. This is because the introduction of noise can cause deformation and interference in the data. Additionally, it is evident that data augmentation yields more pronounced enhancements, particularly for the Camera-PrIMuS dataset, showing an increase of 11.372% in the SeER metric. This dataset encompasses images from real-life scenarios, and the augmented variations better simulate real-world noise and distortions. In contrast, PrIMuS dataset images are typically scanned from printed music scores, exhibiting relatively higher clarity and uniformity, resulting in a comparatively less noticeable impact from these augmentation techniques, with an increase of 5.5525% in the SeER metric. Given that musical notes themselves are small, a significant pixel offset can affect the recognition of individual notes. The ConvNet structure in the CBLSTM method is relatively simple and may not effectively learn from the introduced deviations. Future optimizations to enhance the model's feature capabilities can help reduce symbol error rates.

2) EFFECTIVENESS OF RESIDUAL ConvNet STRUCTURE

The ConvNet in the CBLSTM network was enhanced to create a Residual ConvNet, forming the Residual BLSTM (RBLSTM). The performance of models before and after enhancing the ConvNet under the same experimental conditions was compared. CBLSTM network and RBLSTM were trained separately, and their SyERs were compared.

The changes in loss values during model training for both networks with the number of iterations are illustrated in Figure 13. It is evident that the RCBLSTM network consistently maintains lower loss values than the CBLSTM network with each iteration. After approximately 1,500 training rounds, the RCBLSTM network's loss values have reduced to 5 and stabilized, while the CBLSTM network only decreases to around 10 with substantial fluctuations. This indicates that the Residual ConvNet effectively addresses the non-convergence issue and provides smoother descent in the loss function during training.

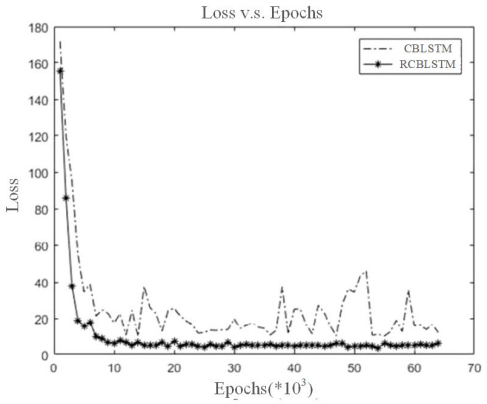

FIGURE 13. Loss values compared with baseline methods.

Furthermore, the SyERs were compared between the two algorithms on the validation set every 1000 iterations, as depicted in Figure 14. Throughout the training process, the SyER for the CBLSTM network initially dropped to around 4%. However, after approximately 52,000 iterations, a significant increase in the SyER is observed, indicating that the model did not converge effectively. In contrast, the RCBLSTM network consistently reduced its SyER to below 2%, with less fluctuation. This demonstrates a noticeable improvement in note recognition accuracy with the Residual ConvNet. Consequently, the Residual ConvNet not only enhances model accuracy but also addresses model degradation issues, improving model generalization capabilities.

3) EFFECTIVENESS OF DEEP MULTILEVEL CASCADE STRUCTURE

To validate the effectiveness of the proposed deep multilevel cascade structure, features from different convolution layers were extracted, including features from C_1, C_3, C_5, and F_4, as shown in Figure 15.

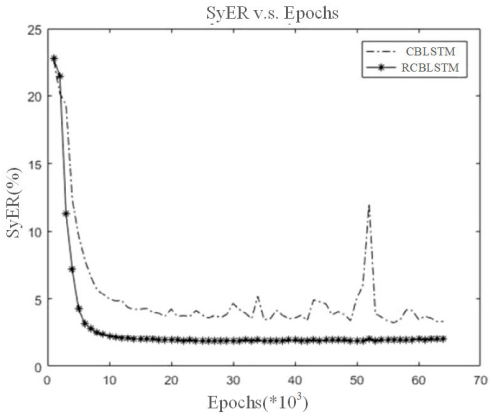

FIGURE 14. SyERs compared with baseline methods.

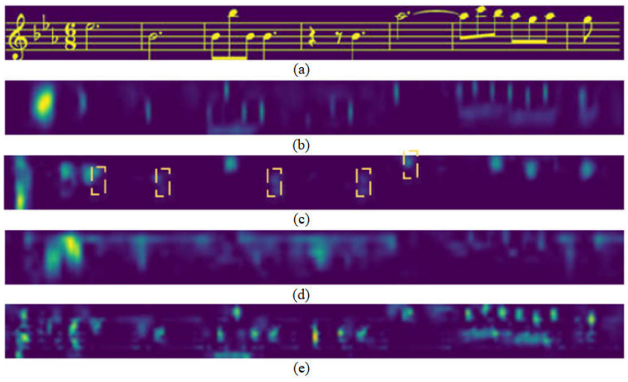

FIGURE 15. Visualization of different layer features. (a) original sheet music. (b) Shallow feature map C_1. (c) Deeper feature map C_3. (d) Deep feature map C_5. (e) Multilevel fusion feature map F_4.

When compared with the original sheet music in Figure 15(a), it can be observed that the feature map C_1 in Figure 15(b), derived from shallow convolution layers, emphasizes the extraction of position information for note elements like stems and beams but captures minimal information about time signatures, note durations, and barlines. In Figure 15(c), feature map C_3 supplements relevant information about dotted notes. It not only includes note positions but also captures information like clefs. Figure 15(d) shows that deep convolution layers no longer extract easily interpretable information but focus on more abstract semantic information, with a stronger emphasis on extracting information about time signatures, rests, and other elements while losing basic note position information. The effect of fusing features from different convolution layers is shown in Figure 15(e), demonstrating that feature map F_4 contains more comprehensive information, confirming the representation capabilities of multilevel features for notes.

Subsequently, multilevel cascade feature fusion was incorporated into the RCBLSTM network, forming the MCRCBLSTM network, to evaluate the impact of multilevel feature fusion on the SyER metric. Figure 16 presents a comparison of symbol error rates between the RCBLSTM network and the MCRCBLSTM network on the validation set. It can be seen that the SyER of the MCRCBLSTM network significantly reduced to below 0.5%, demonstrating the effectiveness of multilevel feature fusion in enhancing the model's ability to extract note features and improve note recognition accuracy.

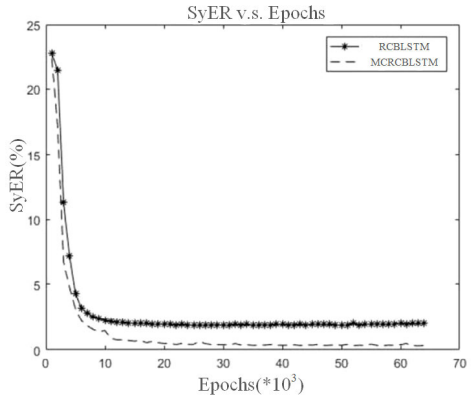

FIGURE 16. SyERs compared with baseline methods about multilevel fusion structure.

4) EFFECTIVENESS OF SIMPLE RECURRENT UNITS

The BiLSTM model in the MCRCBLSTM network was optimized to a BiSRU model further, resulting in the MCRCBSRU (i.e., MCRR) method, which was assessed for convergence speed. Both experiments involved 64,000 iterations. MCRCBLSTM took approximately 16 hours in total, with an average processing time of about 0.92 seconds per iteration, while MCRR required approximately 10 hours in total, averaging about 0.56 seconds per iteration. Figures 17 and 18 presents the results for training loss and symbol error rates on the validation set.

In Figure 17, it can be observed that the MCRCBLSTM achieved a loss of approximately 1.8892 after 1,200 iterations, taking about 18 minutes. In contrast, MCRR reached a loss of approximately 1.5437 after 6,00 iterations, taking about 6 minutes, which is only one-third of the time taken by MCRCBLSTM. Overall, BiSRU model converged significantly faster than the BiLSTM model, approximately three times faster. As shown in Figure 17 and 18, although there was only a 0.08% reduction in SyER, with no significant improvement in accuracy, this result effectively demonstrates the efficiency of the SRU model in parallel time operations.

Finally, the performance of the four ablation models was compared on the same test set. Table 5 shows that the proposed MCRR in this study achieved optimal accuracy in both sequence and symbol error rates, with a SeER

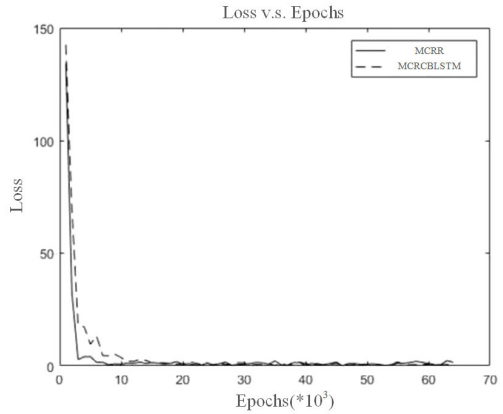

FIGURE 17. Loss values compared with baseline methods about simple recurrent units.

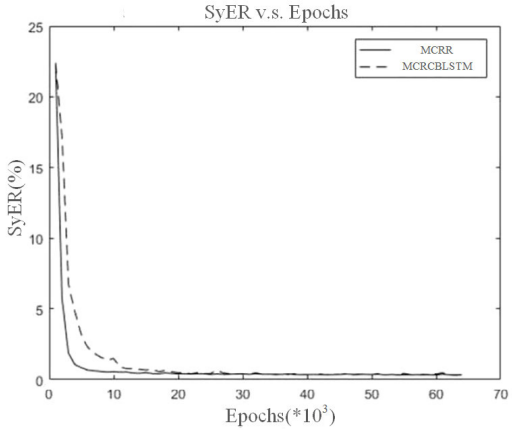

FIGURE 18. SyERs compared with baseline methods about simple recurrent units.

TABLE 5. The comparison performance about ablation models.

Methods	SeER/%	SyER/%
CBLSTM	14.3498	3.2480
RCBLSTM	8.1071	1.8440
MCRCBLSTM	1.4637	0.3312
MCRR	1.4571	0.3234

of approximately 1.4571% and a SyER of approximately 0.3219%. In contrast, the baseline CBLSTM network on the same dataset had a SeER of approximately 14.3498% and a SyER of approximately 3.2480%, which is about ten times less accurate. It is evident that the residual ConvNet and multilevel fusion effectively improved the model's recognition accuracy, validating the effectiveness of the model's optimization.

Considering that the experimental results obtained from the network models are labels for musical notes, they were reconstructed into sheet music images using the Verovio music notation rendering software for a more intuitive comparison as shown in Figure 19.

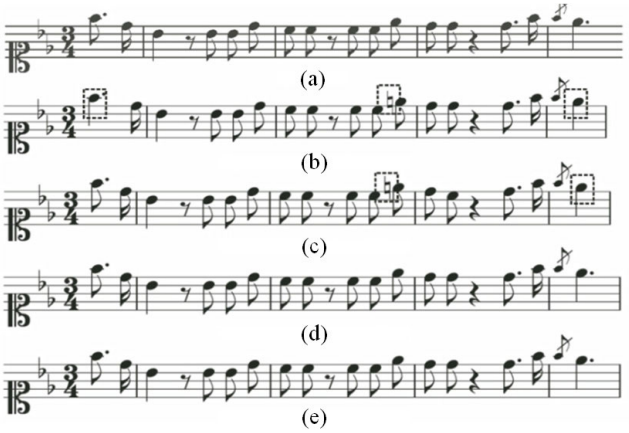

FIGURE 19. Reconstruction results of different ablation models. (a) original sheet music. (b) CBLSTM. (c) RCBLSTM. (d) MCRCBLSTM. (e) MCRR.

Figure 19(a) shows the original music sheet, and the test results using CBLSTM, RCBLSTM, MCRCBLSTM, and MCRR are presented in Figure 19(b) to Figure 19(e). A comparison between Figure 19(b) and Figure 19(c) demonstrates that the inclusion of a residual ConvNet effectively recognized the stems of the notes in the first measure. Figure 19(d) showcases the correction of the RCBLSTM network's erroneous predictions of accidentals in the third measure, while also making correct predictions for dotted notes in the fifth measure. Furthermore, it can be observed from Figure 19(d) to Figure 19(e) that the faster convergence of the model did not negatively impact its accuracy. Through these experiments, MCRR framework significantly improved the recognition of challenging note elements, such as note stems, dotted notes, accidentals, and grace notes. Additionally, our method enhanced the model's training convergence speed, reducing the overall training time.

E. CONTRAST EXPERIMENTS

1) COMPARISON WITH STATE-OF-THE-ART (SOTA) METHODS

In this study, we compared the MF-RC-BiSRU method with other SOTA methods, i.e., ConvNet-STN [10] and DWD [43]. The experimental results are presented in Tables 6 & 7.

Firstly, it can be observed that the proposed MCRR method has achieved the best performance across different evaluation metrics, including SeER, SyER, pitch accuracy, type accuracy, note accuracy, and training time, on both datasets. The ConvNet-STN model is designed specifically for accurate recognition of musical notes without prior knowledge of unexpected events. However, when applied to entire music scores, its feature extraction capability is limited due to unexpected symbols constituting a small portion of the score and the existence of many other types of musical symbols. This limitation leads to decreased accuracy in music score recognition. On the other hand, the DWD model can

TABLE 6. The comparison performance about sota models in the primus dataset.

Methods	SeER /%	SyER /%	Pitch accuracy /%	Type accuracy /%	Note accuracy /%	Training time (s/epoch)
ConvNet-STN	16.8056	5.0208	89.7107	94.0273	85.5620	0.98
DWD	18.5609	8.7811	95.2074	96.0049	91.1255	1.21
MCRR	1.4571	0.3234	97.0000	97.1298	94.4447	0.56

TABLE 7. The comparison performance about sota models in the camera-primus dataset.

Methods	SeER /%	SyER /%	Pitch accuracy /%	Type accuracy /%	Note accuracy /%	Training time (s/epoch)
ConvNet-STN	27.0298	9.3975	70.0000	72.3987	68.3274	2.79
DWD	25.1475	7.9030	82.1123	80.9635	78.5064	3.50
MCRR	5.1488	1.0612	90.7544	91.5520	88.2094	1.93

recognize all types of musical notes but displays varying error rates among different note types. In the PrIMuS dataset, the recognition rate for entire notes is less than 80%, while for sextuplets, it reaches 95%, resulting in a high overall error rate. Additionally, we observed that the average processing time of the DWD method is nearly twice that of the MCRR method. DWD needs to learn the positional information of each note, which increases the model's parameters and training time. In comparison, the ConvNet-STN model is simpler, with fewer parameters, and its training time is similar to MCRR. Regarding the Camera-PrIMuS dataset, it is evident that the performance of all methods has significantly declined compared to the PrIMuS dataset. However, there is a distinct reversal in the status between ConvNet-STN and DWD, indicating that the DWD method has better robustness and can generalize to more complex music sheet images. The advantage of the MCRR method, compared to the PrIMuS dataset, is even greater (except for slight differences in SyER), demonstrating the crucial role played by the proposed data augmentation and multilevel cascade fusion technique in enhancing the robustness and effectiveness of our method. Additionally, it can be observed that the performance of all methods in terms of pitch accuracy, type accuracy, and note accuracy is almost above 80%, which is undoubtedly encouraging. However, the results of SeER and SyER are evidently unsatisfactory. The reasons lie in the assessment metrics: Pitch accuracy evaluates the model's accuracy in recognizing pitch, Type accuracy assesses the correct identification of different symbol types, and Note accuracy focuses on the accuracy of complete attributes of notes including pitch and duration. In contrast, SeER and SyER also consider overall error rates at the sequence and symbol levels, depending not only on individual element accuracy but also on the relationships and sequences between elements. Lastly, concerning training time, the methods in the Camera-PrIMuS dataset have experienced an exponential increase compared to the PrIMuS dataset, as the Camera-PrIMuS dataset contains more complex, larger images from real-life scenarios with increased noise and distortions. Handling this more complex data increases the training time per epoch. In contrast, the PrIMuS dataset is typically simpler and more organized, requiring less time to complete each epoch's training.

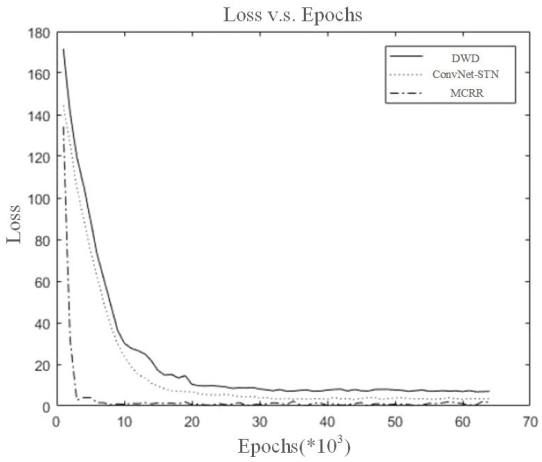

FIGURE 20. Loss values compared with SOTA methods.

Figure 20 shows that the MCRR method's loss value stabilizes after about 600 iterations, while DWD reaches a loss value of around 10 after 2,000 iterations, and ConvNet-STN reaches a loss value of around 3 after 3,000 iterations, indicating slower convergence rates. Therefore, the MCRR framework exhibits good performance in both note recognition accuracy and convergence speed.

2) COMPARISON WITH COMMERCIAL SOFTWARE

We compared the proposed method with results from three commercial optical sheet music recognition software, i.e., Capella-scan, PhotoScore, and SmartScore. The comparison was conducted on 50 music sheets with a total of 1,132 notes, including 107 notes from five enhanced music sheets. As shown in Table 8, our method achieved optimal symbol and sequence error rates, with only five note recognition errors, mainly occurring in two music sheets. Among the three commercial software, Capella-scan had the lowest sequence and symbol error rates, while SmartScore showed

numerous note recognition errors due to issues with barline segmentation.

TABLE 8. The comparison performance about commercial software.

Methods	SeER/%	SyER/%
PhotoScore	42	3.3569
SmartScore	76	46.0247
Capella-scan	36	2.7385
MCRR	4	0.4417

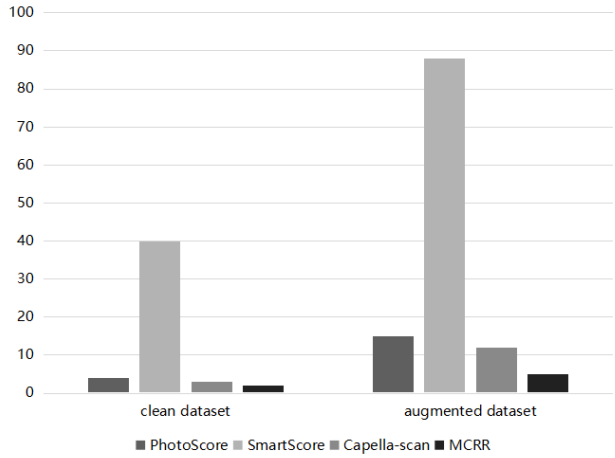

FIGURE 21. The SyER visual comparison about commercial software for data augmentation.

Comparing the four methods on clean music sheet images and those with noisy and deformed music sheet images, as shown in Table 9 and Figure 21, SyERs for commercial software significantly increased on the dataset with noisy and deformed music sheets. Although the proposed method also experienced an increase in error rates on the noisy and deformed music sheet dataset, the increase was relatively small, indicating improvements in the robustness.

TABLE 9. The syer comparison about commercial software for data augmentation (units:%).

Methods	Clean	Augmented
PhotoScore	1.9512	16.8224
SmartScore	41.4634	89.7196
Capella-scan	1.4634	14.9533
MCRR	0.2926	1.8692

In the recognition process, Capella-scan tends to misinterpret rests as numbers and clefs as rests, and these errors are quite concentrated, especially in datasets containing noisy and distorted sheet music images. Additionally, it encounters errors in recognizing beams when eighth notes are connected without any issue to the understanding of the music, but these results contradict common music notation knowledge, as shown in Figure 22(c). On the other hand, Capella-scan

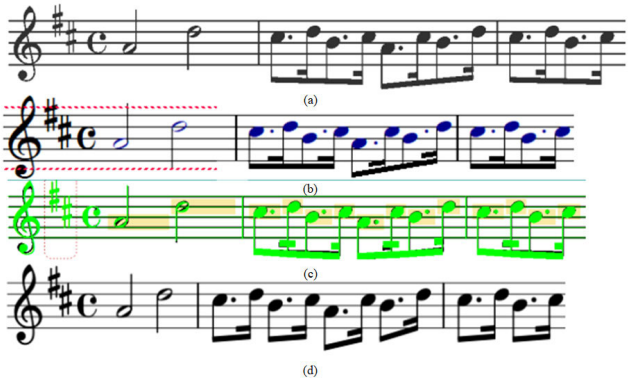

FIGURE 22. Different recognition results for beams when connecting notes of unequal durations. (a) original sheet music. (b) PhotoScore. (c) Capella-scan. (d) MCRR.

excels in accurately recognizing appoggiaturas, ties, and slurs. As shown in Figure 23, when two notes are closely positioned, making it difficult to distinguish between ties and slurs, Capella-scan can make an accurate judgment.

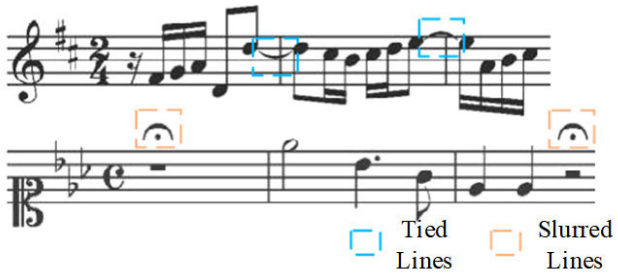

FIGURE 23. Comparison of tied and slurred lines.

PhotoScore's recognition errors are primarily centered around the positions of accidentals, and the relative positions often exhibit frequent deviations, making it difficult to accurately determine the target of the accidentals and sometimes causing them to overlap with surrounding notes. In comparison to Capella-scan, PhotoScore has relatively poor recognition performance for tied and slurred lines. However, PhotoScore excels in accurately recognizing beams, as shown in Figure 22(b). It can also precisely identify notes of unequal durations when they are connected. The accuracy of PhotoScore's recognition is closely related to the density and size of notes in the score. When the score contains a high density of notes, it may lead to significant errors in recognizing simple notes. Conversely, recognition accuracy significantly improves when notes occupy a larger number of pixels.

SmartScore's recognition is highly dependent on the dimensions of the sheet music image. When the image size is small, it might either produce no recognition results or considerably slow down the recognition process, causing the software to become unresponsive. Selecting an appropriate image size requires iterative adjustments, and even when a suitable size is chosen, it tends to have serious issues with barline divisions, which affect note accuracy. However,

when it is employed for the recognition of specific musical instruments, the error rate significantly decreases.

In contrast, the proposed method has minimal recognition errors, affecting only five musical symbols, which are specifically related to time signatures and key signatures. Since our method was not designed to recognize numbers, and each sheet of music in the dataset contains at most one occurrence of numbers, this can lead to significant errors in recognizing time signatures. Regarding key signatures, the differentiation between Bb major and G minor, as well as F major and D minor, is based on the number and position of flat symbols. When the note density in the score increases, errors may occur in recognizing the number and position of flat symbols.

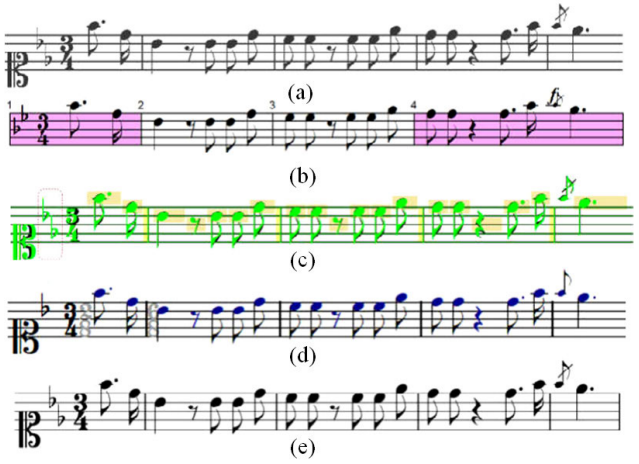

FIGURE 24. Comparison of test results for commercial software. (a) original sheet music. (b) SmartScore (c) Capella-scan. (d) PhotoScore. (e) MCRR.

Comparing the recognition results of the four methods, except for SmartScore, the other three methods exhibit higher accuracy in recognizing note pitches and sextuplets. In contrast, PhotoScore shows significant errors in recognizing the relative positions of accidentals and notes, which can lead to confusion. Additionally, PhotoScore's performance in recognizing ties and slurs is subpar. Capella-scan, despite some inaccuracies in recognizing note beams, outperforms PhotoScore in recognizing ties and slurs. The proposed method demonstrates minimal errors in recognizing challenging elements such as ties, slurs, the relative positions of accidentals and notes, and note beams. However, there are some issues with recognizing specific key signature positions. As shown in Figure 24, in the case of a relatively clean sheet music, Capella-scan correctly recognizes clefs, key signatures, and grace notes, but it makes an error in recognizing the time signature, identifying 3/4 as 5/4. On the other hand, PhotoScore encounters problems in key signature recognition, where it identifies only one flat symbol and has errors in recognizing the tails of grace notes. SmartScore loses information regarding key signatures and exhibits reduced recognition accuracy in sections where bar lines are not accurately divided. Finally, the proposed method does not exhibit any errors in the same sheet music.

From the above experimental results, it is evident that the proposed algorithm achieves lowest symbol error rates, especially in datasets containing noisy and distorted sheet music images. Our method enhances recognition capabilities in the presence of noise, indicating improved robustness. Compared to commercial software, the method achieves a lower symbol error rate and excels in recognizing rests, note beams, ties, and slurs.

V. DISCUSSIONS

A. METHOD SELECTION AND IMPLICATIONS

Based on the attained results and the outlined contributions, the selection of this method for music score recognition can be comprehensively explained:

1) Integration of advanced framework: The proposed optical sheet music recognition framework amalgamates cutting-edge techniques, primarily combining multilevel cascade residual ConvNet with SRU variants of RNN. The framework's architecture encompasses several critical steps. Firstly, dataset augmentation amplifies dataset complexity, enabling enhanced learning capabilities. Secondly, the use of a residual ConvNet in feature extraction tackles gradient vanishing issues, ensuring effective learning by deep networks. Additionally, multilevel cascade feature fusion techniques merge information from diverse convolutional layers, enriching model generalization and feature representation capabilities. Finally, the incorporation of SRU accelerates model convergence, facilitating faster learning.

2) Advanced feature extraction: The introduced feature extraction network structure strategically combines multilevel cascade feature fusion with residual ConvNet. This approach addresses the potential loss of fine-grained details in deep convolutional networks by amalgamating deep and shallower layer feature information. This fusion produces feature maps with enriched information, empowering the model to learn intricate features crucial for improved note recognition. Extensive experiments validate the effectiveness of this combination in reducing symbol error rates and enhancing convergence, resolving critical issues in music score recognition.

3) Note recognition network enhancement: The proposed note recognition network employs SRU coupled with CTC functions. This configuration efficiently handles alignment challenges in sequential data with label information. CTC optimizes path probability to construct notes while SRU, operating independently of previous information outputs, addresses long-range dependency issues. This union of SRU with CTC significantly accelerates convergence speed, nearly tripling the speed of LSTM networks, while maintaining model accuracy, thus demonstrating superior efficiency in note recognition.

4) Rigorous validation and comparative analysis: The performance of the proposed framework undergoes rigorous validation and comparative analysis. Ablation experiments affirm the effectiveness of image preprocessing, residual

ConvNet, multilevel cascade fusion, and SRU modules. Comparative analysis against existing literature and commercial software showcases superior results in terms of symbol error rates, training times, recognition accuracy, and robustness, further solidifying the credibility and efficiency of the proposed method in music score recognition.

Overall, the method's selection for music score recognition is justified by its holistic framework, combining innovative strategies at various stages, robust validation through experimentation, and outperforming existing methodologies and commercial software, establishing its superiority in the realm of optical music recognition.

B. METHODOLOGICAL CONTRIBUTIONS

The proposed optical sheet music recognition framework integrates several innovative components, including multilevel cascade residual ConvNet with SRU. This framework strategically addresses critical challenges in music score recognition. It involves dataset complexity augmentation, gradient vanishing mitigation through residual ConvNet in feature extraction, multilevel cascade feature fusion to enrich feature representation, and the utilization of SRU for accelerated model convergence. The introduction of a feature extraction network that combines multilevel cascade feature fusion with residual ConvNet significantly enhances the model's ability to learn fine-grained features crucial for note recognition. Moreover, the incorporation of SRU with CTC functions improves alignment handling in sequential data, accelerating convergence speed while maintaining accuracy.

C. ADVANTAGES AND DISADVANTAGES

The method's advantages lie in its superior recognition accuracy, robustness across datasets, and improved convergence speed, particularly in symbol recognition within music scores. Notably, its feature extraction strategies effectively preserve fine-grained details, resulting in reduced symbol error rates. However, it should be noted that the method's computational complexity, especially concerning multilevel feature fusion, might require careful optimization for optimal performance across diverse datasets.

D. FUTURE WORKS

The presented method has reduced the recognition error rates for challenging musical symbols. However, there are still instances of recognition errors, particularly in the accurate recognition of certain key signatures. This could be attributed to the scarcity of challenging musical symbols in real sheet music images. Features associated with these symbols may be relatively underrepresented in the dataset, leading to suboptimal learning outcomes for the model. Even with high-dimensional feature models, there may be a risk of overfitting when dealing with infrequent data. Therefore, in scenarios with limited data but a demand for high note recognition rates, it might be beneficial to introduce prior knowledge constraints. For example, combining time signatures, key signatures, and information about notes within each measure with constrained parameter learning to further enhance the model's representation capabilities.

VI. CONCLUSION

Sheet music recognition, as one of the typical technologies for digitizing music information, holds great potential for a wide range of applications, including music content storage, editing, and supporting music education with various instruments. With the widespread use of electronic devices and the growing demand for related software, sheet music recognition is evolving towards recognizing more complex musical scores with higher precision. Conventional sheet music recognition methods based on generic frameworks tend to have complex workflows, lower accuracy, and limited scalability. When new types of musical notations emerge, adapting these methods to recognize the characteristics of new notations becomes challenging. Therefore, in this paper, we propose a deep multilevel cascade residual recurrent (MCRR) framework for sheet music recognition based on the deep learning algorithm. The proposed method primarily consists of three components: image preprocessing, feature extraction, and musical note recognition. Finally, we validate the effectiveness of our model optimization through several comparative experiments and compare it with existing literature and commercial software to highlight the performance of the proposed algorithm.

REFERENCES

[1] A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi, M. Sharifi, N. Zeghidour, and C. Frank, "MusicLM: Generating music from text," 2023, *arXiv:2301.11325*.

[2] L. Carvalho and G. Widmer, "Passage summarization with recurrent models for audio-sheet music retrieval," 2023, *arXiv:2309.12111*.

[3] A. Ríos-Vila, D. Rizo, J. M. Iñesta, and J. Calvo-Zaragoza, "End-to-end optical music recognition for pianoform sheet music," *Int. J. Document Anal. Recognit.*, vol. 26, pp. 1–16, May 2023.

[4] V. Mahipal, S. Ghosh, I. T. Sanusi, R. Ma, J. E. Gonzales, and F. G. Martin, "DoodleIt: A novel tool and approach for teaching how CNNs perform image recognition," in *Proc. 25th Australas. Comput. Educ. Conf.*, Jan. 2023, pp. 31–38.

[5] Q. N. Vo and G. Lee, "Binarization of music score images using line width transform," in *Proc. 21st Korea-Japan Joint Workshop Frontiers Comput. Vis. (FCV)*, Jan. 2015, pp. 1–4.

[6] A. Rebelo and J. S. Cardoso, "Staff line detection and removal in the grayscale domain," in *Proc. 12th Int. Conf. Document Anal. Recognit.*, Aug. 2013, pp. 57–61.

[7] T. Wu, Q. Li, and X. Guan, "Lightweight staff removal method based on multidimensional local binary pattern and XGBoost," *Laser Optoelectron. Prog.*, vol. 56, no. 6, 2019, Art. no. 061006.

[8] J. Calvo-Zaragoza, A.-J. Gallego, and A. Pertusa, "Recognition of handwritten music symbols with convolutional neural codes," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 1, Nov. 2017, pp. 691–696.

[9] J. Hajic Jr., M. Dorfer, G. Widmer, and P. Pecina, "Towards full-pipeline handwritten OMR with musical symbol detection by U-Nets," in *Proc. ISMIR*, 2018, pp. 225–232.

[10] K.-Y. Choi, B. Coüasnon, Y. Ricquebourg, and R. Zanibbi, "Bootstrapping samples of accidentals in dense piano scores for CNN-based detection," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 2, Nov. 2017, pp. 19–20.

[11] J. Calvo-Zaragoza, J. J. Valero-Mas, and A. Pertusa, "End-to-end optical music recognition using neural networks," in *Proc. 18th Int. Soc. Music Inf. Retr. Conf. (ISMIR)*, 2017, pp. 23–27.

[12] L. Tuggener, Y. P. Satyawan, A. Pacha, J. Schmidhuber, and T. Stadelmann, "The DeepScoresV2 dataset and benchmark for music object detection," in *Proc. 25th Int. Conf. Pattern Recognit. (ICPR)*, Jan. 2021, pp. 9188–9195.
[13] J. Calvo-Zaragoza and D. Rizo, "End-to-end neural optical music recognition of monophonic scores," *Appl. Sci.*, vol. 8, no. 4, p. 606, Apr. 2018.
[14] I. Fujinaga, "Optical music recognition using projections," Tech. Rep., 1988.
[15] D. Blostein and H. S. Baird, "A critical survey of music image analysis," in *Structured Document Image Analysis*. Berlin, Germany: Springer, 1992, pp. 405–434.
[16] D. Bainbridge and T. Bell, "The challenge of optical music recognition," in *Computers and the Humanities*, vol. 35, 2001, pp. 95–121.
[17] W. Homenda, "Optical music recognition: The case study of pattern recognition," in *Proc. 4th Int. Conf. Comput. Recognit. Syst. (CORES)*. Berlin, Germany: Springer, 2005, pp. 835–842.
[18] A. Rebelo, G. Capela, and J. S. Cardoso, "Optical recognition of music symbols: A comparative study," *Int. J. Document Anal. Recognit.*, vol. 13, no. 1, pp. 19–31, Mar. 2010.
[19] G. Jones, B. Ong, I. Bruno, and N. G. Kia, "Optical music imaging: Music document digitisation, recognition, evaluation, and restoration," in *Interactive Multimedia Music Technologies*. Hershey, PA, USA: IGI Global, 2008, pp. 50–79.
[20] Q. N. Vo, S. H. Kim, H. J. Yang, and G. Lee, "An MRF model for binarization of music scores with complex background," *Pattern Recognit. Lett.*, vol. 69, pp. 88–95, Jan. 2016.
[21] J. Hatori, T. Matsuzaki, Y. Miyao, and J. Tsujii, "Incremental joint POS tagging and dependency parsing in Chinese," in *Proc. 5th Int. Joint Conf. Natural Lang. Process.*, 2011, pp. 1216–1224.
[22] H.-N. Bui, I.-S. Na, and S.-H. Kim, "Staff line removal using line adjacency graph and staff line skeleton for camera-based printed music scores," in *Proc. 22nd Int. Conf. Pattern Recognit.*, Aug. 2014, pp. 2787–2789.
[23] M. Martin, "On-line support vector machine regression," in *Proc. Eur. Conf. Mach. Learn.* Berlin, Germany: Springer, 2002, pp. 282–294.
[24] A. Konwer, A. K. Bhunia, A. Bhowmick, A. K. Bhunia, P. Banerjee, P. P. Roy, and U. Pal, "Staff line removal using generative adversarial networks," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 1103–1108.
[25] L. Pugin, "Optical music recognitoin of early typographic prints using hidden Markov models," in *Proc. ISMIR*, 2006, pp. 53–56.
[26] W. R. Reed and M. Zhu, "On estimating long-run effects in models with lagged dependent variables," *Econ. Model.*, vol. 64, pp. 302–311, Aug. 2017.
[27] M. V. Mahoney, "Adaptive weighing of context models for lossless data compression," Tech. Rep., 2005.
[28] W. Homenda and T. Sitarek, "Notes on automatic music conversions," in *Proc. Int. Symp. Methodol. Intell. Syst.* Berlin, Germany: Springer, 2011, pp. 533–542.
[29] A. Pacha, K.-Y. Choi, B. Coüasnon, Y. Ricquebourg, R. Zanibbi, and H. Eidenberger, "Handwritten music object detection: Open issues and baseline results," in *Proc. 13th IAPR Int. Workshop Document Anal. Syst. (DAS)*, Apr. 2018, pp. 163–168.
[30] L. Pugin, J. A. Burgoyne, D. Eck, and I. Fujinaga, "Book-adaptive and book-dependent models to accelerate digitization of early music," in *Proc. NIPS Workshop Music, Brain, Cognition*, 2007, pp. 1–8.
[31] G. Taubman, A. Odest, and C. Jenkins, "Musichand: A handwritten music recognition system," Honor thesis, 2005.
[32] A. Rebelo, J. Tkaczuk, R. Sousa, and J. S. Cardoso, "Metric learning for music symbol recognition," in *Proc. 10th Int. Conf. Mach. Learn. Appl. Workshops*, vol. 2, Dec. 2011, pp. 106–111.
[33] D. Bainbridge and T. Bell, "Nineteenth Australasian computer science conference an extensible optical music recognition system," Tech. Rep.
[34] D. Bainbridge and T. Bell, "A music notation construction engine for optical music recognition," *Softw., Pract. Exper.*, vol. 33, no. 2, pp. 173–200, Feb. 2003.
[35] F. Rossant and I. Bloch, "Robust and adaptive OMR system including fuzzy modeling, fusion of musical rules, and possible error detection," *EURASIP J. Adv. Signal Process.*, vol. 2007, no. 1, pp. 1–25, Dec. 2006.
[36] R. M. P. Pereira, C. E. F. Matos, G. B. Junior, J. D. S. de Almeida, and A. C. de Paiva, "A deep approach for handwritten musical symbols recognition," in *Proc. 22nd Brazilian Symp. Multimedia Web*, Brazil, Nov. 2016, pp. 191–194.
[37] W. Zhou and J. Jia, "A learning framework for shape retrieval based on multilayer perceptrons," *Pattern Recognit. Lett.*, vol. 117, pp. 119–130, Jan. 2019.
[38] C. Wen, A. Rebelo, J. Zhang, and J. Cardoso, "A new optical music recognition system based on combined neural network," *Pattern Recognit. Lett.*, vol. 58, pp. 1–7, Jun. 2015.
[39] J. Sober-Mira, J. Calvo-Zaragoza, D. Rizo, and J. M. Iñesta, "Pen-based music document transcription," in *Proc. 14th IAPR Int. Conf. Document Anal. Recognit. (ICDAR)*, vol. 2, Nov. 2017, pp. 21–22.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[41] A. Graves and A. Graves, "Connectionist temporal classification," in *Supervised Sequence Labelling With Recurrent Neural Networks*, 2012, pp. 61–93.
[42] K. Kawakami, "Supervised sequence labelling with recurrent neural networks," Tech. Univ. Munich, Munich, Germany, Tech. Rep., 2008.
[43] Y. Zhang, Z. Huang, Y. Zhang, and K. Ren, "A detector for page-level handwritten music object recognition based on deep learning," *Neural Comput. Appl.*, vol. 35, no. 13, pp. 9773–9787, May 2023.
[44] J. Calvo-Zaragoza and D. Rizo, "Camera-PrIMuS: Neural end-to-end optical music recognition on realistic monophonic scores," in *Proc. ISMIR*, 2018, pp. 248–255.
[45] J. C. Martinez-Sevilla, A. Rios-Vila, F. J. Castellanos, and J. Calvo-Zaragoza, "A holistic approach for aligned music and lyrics transcription," in *Proc. Int. Conf. Document Anal. Recognit.* Cham, Switzerland: Springer, 2023, pp. 185–201.
[46] F. J. Castellanos, J. Calvo-Zaragoza, and J. M. Inesta, "A neural approach for full-page optical music recognition of mensural documents," in *Proc. ISMIR*, 2020, pp. 558–565.
[47] A. Ríos-Vila, J. M. Iñesta, and J. Calvo-Zaragoza, "End-to-end full-page optical music recognition for mensural notation," in *Proc. ISMIR Hybrid Conf.*, 2022.
[48] A. Ríos-Vila, D. Rizo, and J. Calvo-Zaragoza, "Complete optical music recognition via agnostic transcription and machine translation," in *Proc. Int. Conf. Document Anal. Recognit.* Cham, Switzerland: Springer, 2021, pp. 661–675.

PING YU was born in Weifang, Shandong, China, in 1983. She received the bachelor's degree from the Performance Department, Xi'an Conservatory of Music, and the Graduate degree from the School of Music, Shandong Academy of Art. She is currently with the School of Music, Weifang Academy of Art, mainly focusing on vocal singing. Her research interest includes vocal music and related vocal music fields.

HAILING CHEN was born in Hepu, Guangxi. She received the master's degree in musicology from the Guangxi Academy of Arts. Currently, she is a Lecturer with the School of Music and Dance, Heze University.