applied sciences

MDPI

Article

Improved Patch-Mix Transformer and Contrastive Learning Method for Sound Classification in Noisy Environments

Xu Chen [1], Mei Wang [1,2,*], Ruixiang Kan [3] and Hongbing Qiu [3]

[1] College of Computer Science and Engineering, Guilin University of Technology, Guilin 541006, China; 1020221066@glut.edu.cn
[2] College of Physics and Electronic Information Engineering, Guilin University of Technology, Guilin 541006, China
[3] School of Information and Communication, Guilin University of Electronic Technology, Guilin 541006, China; bbklasnic@glut.edu.cn (R.K.); qiuhb@guet.edu.cn (H.Q.)
* Correspondence: mwang@glut.edu.cn

Abstract: In urban environments, noise significantly impacts daily life and presents challenges for Environmental Sound Classification (ESC). The structural influence of urban noise on audio signals complicates feature extraction and audio classification for environmental sound classification methods. To address these challenges, this paper proposes a Contrastive Learning-based Audio Spectrogram Transformer (CL-Transformer) that incorporates a Patch-Mix mechanism and adaptive contrastive learning strategies while simultaneously improving and utilizing adaptive data augmentation techniques for model training. Firstly, a combination of data augmentation techniques is introduced to enrich environmental sounds. Then, the Patch-Mix feature fusion scheme randomly mixes patches of the enhanced and noisy spectrograms during the Transformer's patch embedding. Furthermore, a novel contrastive learning scheme is introduced to quantify loss and improve model performance, synergizing well with the Transformer model. Finally, experiments on the ESC-50 and UrbanSound8K public datasets achieved accuracies of 97.75% and 92.95%, respectively. To simulate the impact of noise in real urban environments, the model is evaluated using the UrbanSound8K dataset with added background noise at different signal-to-noise ratios (SNR). Experimental results demonstrate that the proposed framework performs well in noisy environments.

Keywords: data augmentation; contrastive learning; feature fusion; deep learning; transformer; urban environmental sound recognition

check for updates

Citation: Chen, X.; Wang, M.; Kan, R.; Qiu, H. Improved Patch-Mix Transformer and Contrastive Learning Method for Sound Classification in Noisy Environments. *Appl. Sci.* **2024**, *14*, 9711. https://doi.org/10.3390/app14219711

Academic Editor: Claudio Guarnaccia

Received: 5 September 2024
Revised: 18 October 2024
Accepted: 20 October 2024
Published: 24 October 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Sound serves as an indispensable conduit of information in our daily lives, facilitating human necessary communication, the transmission of factual and emotional content, and the non-verbal expression of emotional and psychological states through phenomena such as crying, laughter, and sighs. It never fails to help humans perceive their environment, identifying the location [1] and motion state [2] of objects. In real scenarios, humans can capture the effective information contained in sound signals to a certain extent by combining specific semantics, context, and personal emotion. In Environmental Sound Classification (ESC), the recognition and classification of sound signals remain challenging. Environmental sounds refer to common sounds in daily life, such as animal vocalizations, indoor environment sounds, and outdoor noises. However, in real life, environmental sounds are often disturbed or jammed by background noise. In urban noise environments, elevated levels of background noise have a huge impact on the recognition of environmental sounds. This is evidenced by the diminished clarity at low signal-to-noise ratio (SNR) conditions, the potential masking of the target sound source [3], and the requirement for extensive annotated datasets [4]. These factors significantly complicate feature extraction, ultimately leading to recognition errors. Furthermore, obtaining high-quality, accurately

annotated audio data presents a significant challenge in real-world scenarios. As a result, this hinders classification tasks [5]. Therefore, improving the classification accuracy of ESC tasks under noise interference is a challenging research topic and emerges as a pressing research endeavor.

This paper focuses on several key issues in ESC tasks, such as the representation of sound features, the accuracy of classification models, and relevant data augmentation schemes. By addressing these critical challenges, our core research is to focus on developing robust and accurate ESC systems that can operate effectively in noisy environments. The standpoint of our main work stems from addressing urban high-noise environments, meticulously unfolding its central thesis through the adept handling of these intricate challenges. The diverse modules synergize seamlessly, fostering the development of a robust, adaptable, and precise identification system. The innovative Patch-Mix feature fusion strategy and the prowess of contrastive learning, our groundbreaking Contrastive Learning-based Audio Spectrogram Transformer (CL-Transformer) framework, attain remarkable performance in the sound classification method. Our proposed approach is supposed to offer a dependable and precise solution tailored for the real-world situation, thereby enhancing efficacy and accuracy in these demanding environments. A detailed description of the core methods of this paper will be presented in the related work section.

1.1. Related Works

Unlike musical and speech audio signals, urban environmental sounds consist of indoor and outdoor conditions, which contain random and unpredictable background noise. This results in more complex characteristics of environmental sound information, making it a primary issue in the ESC field to extract such features and reduce the impact of noise on model recognition accuracy.

In the existing research on audio signals, traditional feature extraction methods often focus on time domain information, spatial information, frequency domain characteristics, etc. In recent years, feature extraction approaches combining time frequency characteristics with statistical characteristics have played a significant role, such as Mel Spectrogram, Log-Mel Spectrogram, Mel-Frequency Cepstral Coefficients (MFCC), and Gammatone Frequency Cepstral Coefficients (GFCC) [6]. Due to different extraction methods, these acoustic features have unique application scenarios and advantages [7]. With the development of deep learning, many researchers have transcended the limitations of relying solely on individual acoustic features [8], instead adopting feature fusion to integrate multiple acoustic features to harness their collective strengths [9]. For example, [10] analyzed various acoustic features and experimentally studied whether different feature fusions affect classification results. Experiments demonstrate that some feature fusions, such as MFCC and Chroma STFT, can enhance classification model accuracy. In [11], logarithmic operations on Log-Mel spectrograms produced Log2Mel (L2M) and Log3Mel (L3M). The fusion of Log-Mel with L2M and L3M results in new representations, making it easier for these feature sets to capture detailed texture characteristics and abstract features in ESC tasks and, thus, achieves advanced results in recognition methods based on this. In [12], experiments show that duplicating the Log-Mel spectrogram into three channels as input features yielded higher classification accuracy compared to a single Log-Mel spectrogram. When single features cannot sufficiently represent signals, using multi-feature fusion to jointly represent signals is an effective solution. However, at the current research stage, whether using splicing techniques [11] or dimensional stacking [12] to fuse features, it undoubtedly significantly increases the dimensionality of feature vectors. This approach underscores the potential of exploiting feature diversity and redundancy to refine model performance in complex urban environments. Therefore, it is necessary to seek new solutions on this basis.

In order to represent acoustic signals, the proper and clever selection of classification models in ESC tasks goes hand in hand with representing acoustic signals effectively. In [13], a sub-spectrogram segmentation with score level fusion framework for ESC is proposed, along with a Convolutional Recurrent Neural Network (CRNN) to improve

classification accuracy. To tackle challenges such as vanishing gradients, high network loss, and low accuracy, [14] propose Long Short-Term Memory Neural Networks (LSTM) and Gated Recurrent Unit (GRU) neural networks, constructing a deep recurrent neural network model. However, recognizing environmental sounds in noisy scenarios typically poses a unique challenge due to the sheer diversity and often imbalanced nature of sound data. Some categories of data may be relatively abundant, while others are very scarce. Through transfer learning, pre-trained models on data-rich categories can be transferred to data-scarce categories and, thus, improves overall data utilization. This can significantly reduce the computational resources needed to train models from scratch. Furthermore, the direct training of core models on limited datasets in ESC tasks can easily lead to overfitting. Transfer learning can enhance the recognition effect of the core methods in this paper by leveraging models pre-trained on extensive datasets. This approach not only strengthens the core methods discussed in our manuscript but also addresses overfitting concerns, thereby enhancing the overall recognition effectiveness indeed. More importantly, the application of transfer learning has the potential to improve the performance of ESC models. By pre-training on ESC tasks, the core recognition model can obtain better initial parameters and feature representations. It helps algorithms converge quickly and achieve optimal solutions within the hyperplane, resulting in better performance on the validation set. Therefore, transfer learning is becoming increasingly important in the study of environmental sound classification [15]. Some conventional computer vision methods are being expanded. Network architectures tailored for image analysis have gradually ventured into ESC applications [14,16]. However, current ESC task methods using pre-trained Convolutional Neural Networks (CNNs) often resort to redefining the last layer to address sound classification challenges [11]. A notable limitation of this approach lies in its inability to capture the intricate long-term dependency relationships in audio signals [17], and since most of these models are transferred from the image domain, they focus more on the physical meaning and distribution characteristics of images, requiring higher training weights for parts that can effectively reflect abstract feature distribution characteristics and semantic cues. Thus, merely fine-tuning the last output layer and directly applying it to ESC tasks may not yield good results. Consequently, attention mechanisms, as representative optimization strategies, are progressively garnering more and more attentions all over the world. These mechanisms offer a promising avenue to address the limitations of traditional CNN-based approaches, enhancing the model's capability to exploit critical temporal and contextual information within audio signals in the real-world situation.

As attention mechanisms are introduced into the field of sound classification, some research has improved pre-trained network models by seamlessly integrating attention mechanisms [7]. In [3], a ResNet model based on a multi-feature attention mechanism is proposed, which added a Net50_SE attention module to suppress the impact of environmental noise. By redistributing the training weights of the sound signals, this improvement has achieved higher recognition accuracy in noisy environments. Experimental evidence underscores the adaptability of certain attention modules, which can cater to the fine-tuning demands of transfer learning by dynamically adjusting parameter weights. This adaptability empowers the model to interpret increasingly intricate audio signals. Consequently, many previous studies are exploring attention-based solutions [18], with models based on Transformer concepts demonstrating strong feature learning capabilities [19]. The Audio Spectrogram Transformer (AST) model is proposed in [20], and experiments show that AST pre-trained on ImageNet and AudioSet datasets have achieved advanced results in various downstream audio tasks. Using pre-trained models can achieve better generalization on smaller datasets, helping the model obtain better initial parameters and improve training efficiency. However, due to factors such as noise and dataset size, pre-trained models in ESC tasks are also prone to overfitting. Researchers have endeavored to address overfitting through various strategies, including parameter reduction [21] and data augmentation [12]. These methods can only partially reduce overfitting and cannot completely eliminate this issue, which significantly constrains the development of urban environmental sound

recognition algorithms, especially in high-noise environments where this disadvantage is amplified. Therefore, solving the overfitting of pre-trained models in downstream tasks is also crucial.

At present, the research on urban environment sound recognition tasks in flexible and noisy environments remains untouched. Therefore, transferring network models from the image domain to ESC tasks can effectively improve classification accuracy. However, for ESC tasks with smaller datasets, using pre-trained models through transfer learning can accelerate overfitting, which is detrimental to model training [6]. As a key technology in ESC tasks, data augmentation emerges as a widely embraced strategy to mitigate these issues. In [22], a set of offline data augmentation techniques is employed, targeting audio waveforms, performing necessary time-stretching and pitch-shifting. They also incorporate background noise from urban environments to enhance robustness in real-world scenarios. In [23], time warping based on spectrogram augmentation is utilized, randomly stretching or compressing spectrograms through interpolation techniques. Experiments have shown that data augmentation can alleviate overfitting to some extent. Additionally, some researchers use contrastive learning as a regularization scheme to address overfitting. For instance, [12] introduce contrastive learning loss into a ResNet-50-based model to solve this problem, achieving good results. Given the current lack of urban environmental sound datasets under noisy conditions, it is essential to use data augmentation techniques and related regularization schemes to assist model training.

1.2. Contributions

In real life, urban environmental sounds include background noise with low SNR [24]. Therefore, minimizing the impact of noise on classification models and improving the robustness of these models to noise is a crucial issue in ESC tasks. Conventional classification models cannot fully extract audio features, and introducing transfer learning can enhance the specificity of the training process. However, this approach objectively brings the risk of overfitting during training [25]. Therefore, based on existing research schemes, this paper proposes new solutions and improvements.

Acknowledging these limitations, grounded in the current research landscape, we dedicate ourselves to present novel solutions and advancements. We endeavor to address the aforementioned challenges by introducing innovative strategies that not only lessen the influence of noise but promote the generalizability and robustness of ESC models as well, thereby advancing the state-of-the-art in our manuscript. Our main contributions are divided into the following three parts:

(1) A novel Patch-Mix feature fusion scheme is proposed. Unlike traditional methods that perform feature fusion by simply stacking multiple dimensions, the Patch-Mix strategy randomly assembles patches within samples instead of merely merging multiple dimensions. This provides the model with more comprehensive feature information without increasing the feature dimensions. Additionally, introducing patches from noisy samples for replacement can enhance the robustness of the feature learning process in noisy environments, helping the model better adapt to real-world scenarios during training.

(2) The CL-Transformer model is proposed to address the overfitting weaknesses of pre-trained Transformer models during training, especially on small-scale datasets, by introducing a contrastive learning strategy. This approach incorporates contrastive learning loss, providing an effective regularization mechanism for the Transformer model during training, thereby alleviating overfitting issues. Additionally, the contrastive learning mechanism enhances the model's robustness in noisy environments. This strategy aims to improve the generalization ability and performance of the model, and to mitigate the overfitting problems that often occur in downstream tasks after transfer learning. As a result, our novel CL-Transformer model in our manuscript emerges as a more versatile and robust solution, better suited for diverse downstream applications and conditions.

(3) An improved data augmentation technique has been developed to help the model better adapt to different input variations and noise by augmenting the original audio waveforms and spectrograms. This enhances the model's robustness and alleviates overfitting issues in downstream tasks with limited datasets.

In total, our state-of-the-art methods presented in this manuscript bolster the performance and applicability across a broader spectrum of real-world scenarios.

2. Proposed Methods

2.1. Acoustic Feature Extraction

In the field of audio classification, effectively extracting features that are rich in information and robust against noise and interference is crucial for improving the performance and generalizability of models [26]. To overcome the negative impact on environmental sound recognition in urban noisy environments, this paper proposes an innovative feature fusion representation strategy by combining the Patch-Mix feature fusion scheme to reduce the influence of noise and fully represent audio features. It aims to mitigate the influence of noise while comprehensively capturing the key audio features, thereby achieving outstanding and fabulous accomplishments.

Introducing background noise during model training can help the model better adapt to various environments and noises [27]. The statistical characteristics of Gaussian distribution white noise are similar to those of background noise in many natural environments. Therefore, additive white Gaussian noise (AWGN) is often chosen as a high-quality option for simulating environmental background noise in this domain [27]. First, AWGN is incorporated into the original audio to serve as background noise. Secondly, Log-Mel spectrograms are extracted from both the original and the noise-added audio. The Log-Mel spectrograms are used as input for the Transformer model, which can be conceptualized as a temporal sequence, where each temporal instant encapsulates key features spanning diverse frequency dimensions [28]. The Transformer model enhances the Log-Mel spectrogram by incorporating positional information into the time dimension. It is quite convinced that this structure dynamically reallocates weights on a large scale to capture global information and long-term dependencies within the audio signal based on the adaptive attention mechanism. Therefore, using Log-Mel spectrograms as input for the Transformer model fully leverages the advantages of both, enabling efficient processing of audio signals. Figure 1 shows the extraction process of the Log-Mel spectrogram.

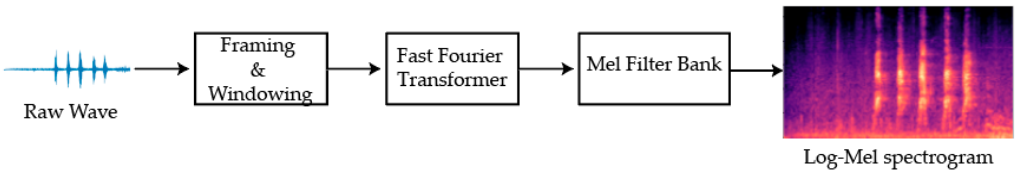


Figure 1. Log-Mel Spectrogram Extraction Process Diagram.

Finally, the spectrogram is divided into multiple patches through Patch Embedding, and then Patch-Mix is used to implement patch-level fusion. Figure 2 shows the overall feature extraction process.

Patch-Mix Feature Fusion

In the existing field of audio classification, feature fusion schemes predominantly rely on multi-scale, multi-channel, and multi-spectral approaches to facilitate fusion, thereby implementing adaptive weighting mechanisms and robust recognition for specific scenarios [29]. However, these existing fusion mechanisms are gradually revealing their limitations in urban environmental sound recognition. Unavoidable noise, being a pervasive factor, can obscure audio signals to varying degrees and potentially overshadow the target audio, making current multi-dimensional fusion techniques insufficient in handling noise interference [3]. Additionally, introducing more parameters and complexity increases

the model fitness to the training data and, thus, raises the risk of overfitting. This implies that the aforementioned methods are no longer suitable for the application scenarios of this study. Therefore, this paper proposes a novel feature fusion scheme to address the feature fusion problem.

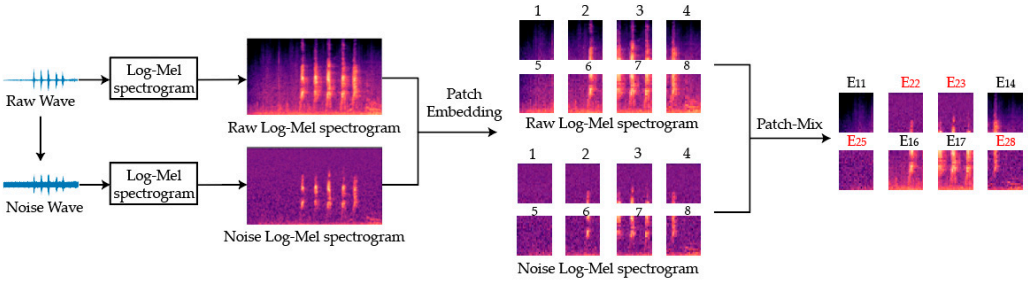


Figure 2. Acoustic Feature Extraction Process Diagram. E_{ij} represents the embedding of the j-th patch of the i-th spectrogram.

First, a random value, λ, is generated using the Beta distribution to control the mixing process between the two audio features during the feature fusion process [29], determining the weight between the two audio features. The Beta function is defined as follows:

$$B(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt \quad (1)$$

where α and β are the two shape parameters of the Beta distribution, which determine the shape of the Beta distribution. When both α and β are 1, the Beta distribution degenerates into a uniform distribution, where all possible values have an equal probability density. This ensures that the generated mixed features effectively reflect the overall information of the feature sets during the feature fusion process, preventing the model from overfitting to a specific spectrogram feature.

Then, the spectrogram is divided into multiple patches through the Transformer's Patch Embedding module, followed by a strategic random mixing process of these patches using the Patch-Mix method. During the mixing process, patches at the same position are swapped without altering their positional encoding function. Subsequently, related simulation experiments are conducted on the baseline model. The specific experimental results, shown in Table 1, demonstrate that this feature fusion technique significantly reduces GPU capacity and inference time compared to traditional feature fusion schemes, and it can represent environmental sounds.

Table 1. Experimental analysis of Patch-Mix feature fusion on the UrbanSound8K dataset.

Feature Fusion Scheme	Input Feature	GPU Capacity (MiB)	Inference Time (ms)	Accuracy (%)
-	Fbank	5866	304	88.68
Dimensional Merge	Fbank, Noise-Fbank	17,284	1240	89.60
Patch-Mix	Fbank, Noise-Fbank	5868	323	90.56

2.2. Combined Data Augmentation Scheme

In ESC tasks, the performance of a model is often constrained by the size of the available dataset. Particularly in cases where the dataset is small, the model is more prone to overfitting, which significantly reduces its performance on unpredictable data [30]. Data augmentation techniques become especially important in this context. By employing reasonable data augmentation schemes tailored to the actual needs of the research scenario, various environments and conditions in the real world can be simulated, providing the model with more diversified training data and enhancing the model's generalization capability [31].

For the research scenario of this paper, a combined audio signal enhancement technique is proposed. First, an offline data augmentation technique for audio waveforms is designed. Initially, silent parts at the beginning and end of the audio are removed. Then, offline data augmentation is performed on the audio by first applying random scaling to simulate time-stretching and pitch-shifting. Next, Dynamic Range Compression (DRC) is used to generate audio samples with different dynamic range characteristics. Last but not least, to standardize the audio input lengths, we implement random padding, setting the UrbanSound8K audio segments to 176,400 samples for 4 s clips and ESC-50 segments to 220,500 samples for 5 s clips, ensuring consistency across the training process.

The augmented audio undergoes Log-Mel spectrogram extraction, followed by spectrogram augmentation techniques such as time warping, time masking, and frequency masking. In the data augmentation process for audio signals, to better simulate the variations of urban environmental sounds and to prevent the model from overfitting to a single type of data transformation, a certain degree of randomness is introduced. This ensures that the same audio signal can receive different data augmentation effects in different training batches. Considering the impact of random masking parameter settings on the experiment, the masked segments on the time axis (t) are set to 1, and the masked segments on the frequency axis (f) are set to 2, with the maximum segment width $T = 32$ and $F = 32$. The specific data augmentation process is shown in Figure 3.

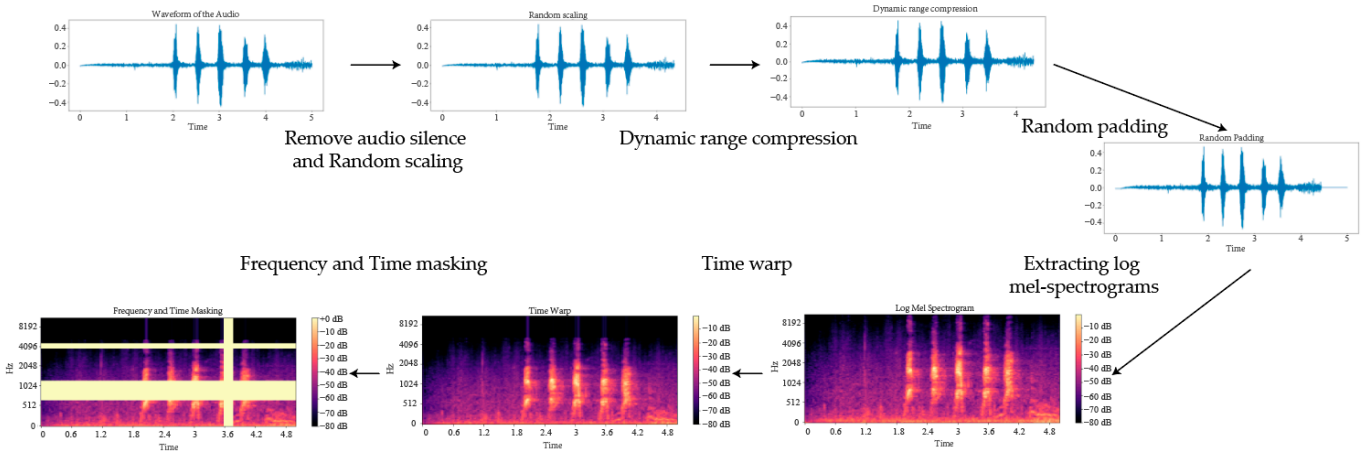


Figure 3. Combined Data Augmentation Process Diagram.

2.3. CL-Transformer Classification Model Based on Contrastive Learning

In most supervised learning methods for audio classification, training is guided by cross-entropy loss. The cross-entropy loss measures the difference between the output probability distribution and one-hot encoded label and the model tries to reduce this loss by mapping the samples of the same class to their respective label vector. The core idea of contrastive loss is to optimize by defining positive and negative pairs, drawing closer samples belonging to the same class (positive pairs) and repelling those from different classes (negative pairs) [12]. In supervised contrastive loss, actual labels are used to generate positive pairs and negative pairs. Introducing supervised contrastive loss as a supplement to cross-entropy loss can further distinguish samples from different classes. By contrasting positive and negative sample pairs, the model is encouraged to learn the subtle differences between different classes, rather than simply memorizing specific labels. Furthermore, this mechanism implicitly mitigates the influence of noise in audio samples, enhancing the overall robustness of the classification system.

Audio Spectrogram Transformer (AST) is a fully attention-based model capable of handling variable-length inputs [32]. Pre-trained AST models demonstrate superior performance in downstream tasks. However, when confronted with a limited dataset for such downstream tasks, the pre-trained model can easily overfit the characteristics of the training data, making it less adaptable to new tasks. As a flexible and strategic regu-

larization technique, the incorporation of contrastive learning loss acts as a formidable safeguard, effectively mitigating the overfitting phenomenon and enhancing the model's generalizability.

Based on the above, this paper proposes an improved CL-Transformer classification model for sound classification in noisy environments. This model framework modifies the Patch Embedding layer of the AST model and introduces the Patch-Mix method for feature fusion. Unlike traditional feature fusion methods that increase the feature space dimensions, Patch-Mix ingeniously offers a more holistic feature representation without augmenting the feature dimensions, simultaneously bolstering resistance capability against noise disturbances. Additionally, supervised contrastive loss is introduced as a supplement to cross-entropy loss. The model is shown in Figure 4.

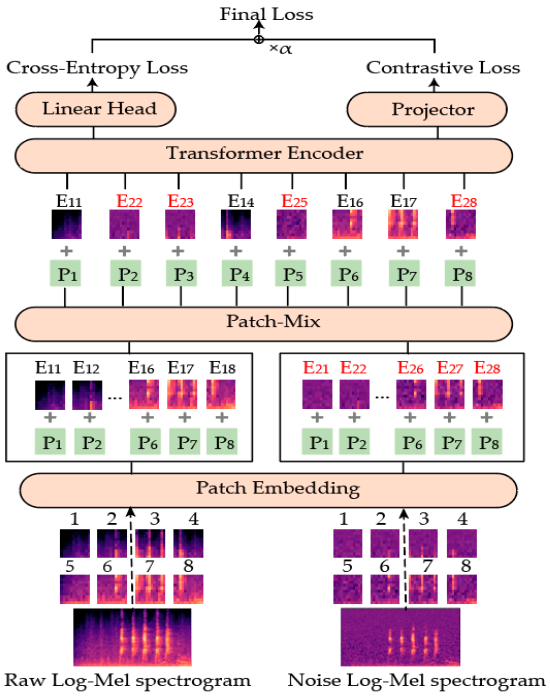


Figure 4. CL-Transformer Network Model Diagram.

First, the model executes Patch Embedding on the input spectrogram, dividing it into a series of patches and adding learnable positional encodings to each patch. Then, Patch-Mix operation is carried out, mixing patches from the original spectrogram and the noisy spectrogram. These mixed patches are used as inputs to the Transformer encoder. Drawing upon the encoder's refined outputs, the model concurrently calculates both cross-entropy loss and contrastive loss, each uniquely encapsulating disparate information of the learning objective. Judging by a strategic redistribution and harmonious fusion of these two loss functions, we derive the definitive final loss, which subsequently steers the training trajectory of our method with superb accuracy.

The baseline model AST is denoted as $f(\cdot)$. The classifier model Linear Head consists of a normalization layer and a fully connected layer. The normalization layer processes the output of the Transformer Encoder, and the fully connected layer connects to the input features to the output label dimensions. The cross-entropy loss function is used to calculate the loss. The loss calculation is as follows:

$$L_i^{cross-entropy} = -\sum_i y_i^{one-hot} \log\left(y_i^{logits}\right) \tag{2}$$

where $y_i^{one-hot}$ is the one-hot encoded true label of sample i, and y_i^{logits} is the classifier output for that sample.

Contrastive loss maximizes the correlation degree between positive pairs of samples and reduces the correlation degree between negative pairs, thereby highlighting the adaptive weight of key sample pairs to be recognized. In the formulation of sample pairs, entities within the same batch and sharing an identical label are rearranged as positive parts, whereas all disparate instances are designated as negative parts relative in our method. The output of the Transformer Encoder layer has a high dimension, and directly applying contrastive loss is not supposed to effectively separate samples of different classes. Directly reducing the output dimension of the Transformer Encoder layer can lead to information loss. To prevent this and compensate for the loss of spatial features, a Projector network layer is deployed as a transitional bridge in time, reducing the dimensionality of the Encoder's output while preserving the essence of the data indeed. The Projector, represented as $g(\cdot)$, is a linear fully connected layer, and its input is the normalized output of the Transformer Encoder layer. Given labels $Y = \{y_1, y_2, \ldots, y_n\}$ and corresponding samples $X = \{x_1, x_2, \ldots, x_n\}$, sample x_i is mapped to the representation vector h_i, where $h_i = f(x_i) \in \mathbb{R}^{D_f}$. The representation vector is reduced in dimension through the projection network $g(\cdot)$, yielding $z(h_i) = g(h_i) \in \mathbb{R}^{D_g}$, where $D_g \ll D_f$. The supervised contrastive loss is then calculated as follows:

$$L_i^{sup-cont} = -\frac{1}{N_{y_i}} log \frac{\sum_{j=1}^{N} \chi_{[y_i=y_j]} \exp\left(sim\left(z_i, z_j\right)/\tau\right)}{\sum_{k=1}^{N} \chi_{[k\neq i]} \exp(sim(z_i, z_k)/\tau)} \tag{3}$$

where N_{y_i} is the number of samples with the same label y_i in the same batch, $\chi_{[y_i=y_j]} \in \{0, 1\}$ evaluates to 1 if and only if $y_i = y_j$, N is the batch size, τ is the temperature parameter used to adjust the scale of the similarity measure between sample pairs, and $sim(\boldsymbol{u}, \boldsymbol{v})$ is the function that calculates the cosine similarity between vectors $\boldsymbol{u}$ and $\boldsymbol{v}$. The calculation formula is:

$$sim(\boldsymbol{u}, \boldsymbol{v}) = \frac{\boldsymbol{u}^T \boldsymbol{v}}{\|\boldsymbol{u}\| \|\boldsymbol{v}\|} \tag{4}$$

The cross-entropy loss function provides global supervision by directly optimizing classification results, ensuring that the model has good classification ability. The supervised contrastive loss function provides local supervision by optimizing the similarity between sample pairs, enabling the model to learn more discriminative feature representations. When integrated harmoniously, these two core loss functions operate together, mutually reinforcing their respective strengths and the robustness of the model. The computation of this hybrid loss, which encapsulates the best of both worlds, is formulated as follows:

$$L_i = \alpha L_i^{sup-cont} + (1 - \alpha) L_i^{cross-entropy} \tag{5}$$

where α is a hyperparameter used to adjust the weight between cross-entropy loss and supervised contrastive loss. Additionally, this paper explores the impact of different α values on the mixed loss through experiments. The relevant experimental results are shown in Table 2.

Table 2. The impact of different α values on the mixed loss.

α	ESC-50 (%)	UrbanSound8K (%)
0.01	95.00	89.85
0.25	95.50	90.92
0.50	96.75	91.39
0.75	95.25	90.80
0.99	94.75	89.73

Table 2 shows that the best results are achieved when the cross-entropy and contrastive loss contribute equally to the loss calculation. We determined the optimal value of α through experiments. Therefore, in the subsequent experiments, we set the value of α to 0.5.

3. Experiment Setup

3.1. Public Environment Sound Datasets

In our manuscript, we have strategically chosen the widely accessible ESC-50 [33] and UrbanSound8K [34] datasets as the standpoints for training process and evaluating our model. These two datasets are publicly available and can be used for sound classification tasks, providing different sound samples suitable for training and evaluating various sound classification models, along with their corresponding ground truth labels.

The ESC-50 dataset contains 2000 five-second sound clips from different environments. These samples are divided into five major categories: animal sounds, natural soundscapes and water sounds, non-speech human sounds, indoor/domestic sounds, and outdoor/urban noises. These 5 major categories are further subdivided into 50 subcategories, including typical urban environment sounds such as car horns, city noise, sirens, and helicopter sounds. The ESC-50 dataset is formally divided into five folders, each containing an equal number of sound samples, and these samples are balanced by type. The label set of ESC-50 is defined as $label = \{0, 1, 2, \ldots, 49\}$, where each element corresponds to a distinct category.

The UrbanSound8K dataset contains 8732 labeled urban sound samples from 10 categories, with each sound sample no longer than 4 s. The 10 categories include air conditioner noise, car horn, children playing, dog barking, drilling, engine idling, gunshot, jackhammer, siren, and street music. While the UrbanSound8K dataset is vividly divided into 10 dedicated folders, it is noteworthy that the distribution of sound samples within each folder is not uniformly balanced in terms of both category representation and quantity. So that it brings us a more realistic and challenging condition for the training process and evaluation classification models. The label set of UrbanSound8K is defined as $label = \{0, 1, 2, \ldots, 9\}$, with each element corresponding to a distinct category.

In the experiments, all audio datasets are resampled at a 44.1 kHz sampling rate. Since the audio samples in the UrbanSound8K dataset have uneven lengths due to the sound clipping process and inconsistencies in the channels used for sound signal collection, the audio length is standardized to 4 s during data preprocessing. For stereo channels, the Patch-Mix method is used to obtain a unique representation.

3.2. Feature Extraction

During the extraction of the Log-Mel spectrograms, background noise is mixed among the audio. Then, both the original audio and the corresponding noise-added audio waveforms are converted into 128-dimensional log-Mel filter bank (Fbank) feature sequences [20], with window and overlap sizes set to 25 ms and 10 ms, respectively. The Patch-Mix operation is applied on the obtained original audio Log-Mel spectrograms and the corresponding noise-added Log-Mel spectrograms to obtain the final mixed features.

3.3. Transfer Learning

Compared to previous audio classification tasks that relied entirely on convolutional models, the AST model, grounded in the self-attention mechanism, demonstrates a remarkable ability to delve deeper into capturing long-range dependencies across both temporal and frequency dimensions in our method. Therefore, this paper adopts the AST model to accomplish the audio classification task. Through experiments, it is found that the AST model exhibited severe overfitting during fine-tuning. To bridge this huge gap, we introduce the novel CL-Transformer model framework, a refinement built upon the adaptive AST model. The CL-Transformer model framework implements the Patch-Mix feature fusion scheme based on Patch Embedding and includes reweighting of the cross-entropy loss and supervised contrastive loss to jointly aid in model training. These updating strategies and key techniques collaborate to cope with overfitting and propel the model towards more effective and robust audio classification functions.

4. Results and Analysis

The AST model is chosen as the baseline model in our manuscript, with the input features being the Fbank features of the original audio. During training, the cross-entropy loss function is used to train the model. The model employs the AdamW optimizer, with a base learning rate of 1×10^{-5} for the ESC-50 and UrbanSound8K datasets. Additionally, the training process is optimized using exponential learning rate decay with a factor of 0.98 and a warm-up period of 10 epochs. The batch size is set to 16. The pre-trained AST model is fine-tuned for 50 epochs.

The experimental environment includes an Intel Core i5-13490F processor (Intel Corporation, Santa Clara, CA, USA) and an RTX 4060 Ti (16 GB) graphics card (NVIDIA, Santa Clara, CA, USA). All experiments are conducted in a Python 3.11.5 and Pytorch 2.1.2 platform.

4.1. The Impact of Transfer Learning on Model Performance

The baseline model AST is initially used for sound classification in the AudioSet dataset and then applied to downstream tasks through transfer learning. Therefore, it is necessary to verify whether transfer learning affects the model's performance. The pre-trained AST model is fine-tuned for 50 epochs, while the non-pre-trained AST model is trained from scratch for 200 epochs. Exponential learning rate decay and a warm-up period are used to optimize the training process, with an initial learning rate of 1×10^{-5} for the pre-trained AST and 1×10^{-4} for the non-pre-trained AST. The experimental results are shown in Table 3.

Table 3. The impact of transfer learning on classification performance. IN represents the ImageNet dataset, and AS represents the AudioSet dataset.

Model Architecture	Pretrain	ESC-50 (%)	UrbanSound8K (%)
AST	-	65.00	79.93
AST	IN	88.00	85.90
AST	IN + AS	92.25	88.68

Compared to the UrbanSound8K dataset, the ESC-50 dataset contains only 2000 audio samples but has 50 categories, making it difficult for a non-pre-trained AST model to achieve high classification accuracy on the ESC-50 dataset. In contrast, it can achieve a relatively high classification accuracy on the UrbanSound8K dataset, which has only 10 categories. Notably, the pre-trained AST model drastically elevates classification accuracy across both smaller and more diverse datasets, underscoring the inadequacy of relying solely on non-pre-trained models for urban environmental sound classification tasks. However, compared to the AST model pre-trained only on the ImageNet dataset [35], the AST model additionally trained on the AudioSet audio dataset performs better. As Table 3 vividly illustrates mentioned above, harnessing transfer learning by applying this pre-trained model to ESC tasks significantly amplifies the accuracy and capability. It also emphasizes the transformative impact of our novel pre-training and transfer learning methods in enhancing model performance for audio classification tasks.

4.2. The Impact of Combined Data Augmentation Scheme on Model Performance

In cases where the dataset is small, data augmentation can significantly improve the model's performance and generalization ability. However, different data augmentation schemes may bring different impacts on these datasets. Hence, we evaluate and compare different combinations of data augmentation schemes. The specific outcomes of these experiments are precisely presented in Table 4.

First, we validated the traditional data augmentation schemes commonly used by many previous studies, which include time stretching and pitch shifting based on audio, as well as time masking and frequency masking based on spectrograms. Then, we add Dynamic Range Compression (DRC) and Time Warping (TW) to these schemes and explored

the impact of different degrees of warping on the experimental results. The experimental results show that traditional data augmentation schemes can significantly improve model performance on the ESC-50 dataset, which has fewer data and more categories. This experiment has demonstrated that our combined data augmentation scheme generates more diverse audio variants, which helps the model learn more useful feature information to improve performance. When compared against scenarios where no data augmentation is employed, our proposed combined data augmentation scheme exhibits a remarkable improvement of 3.25% on the ESC-50 dataset and a commendable 2.24% gain on the UrbanSound8K dataset. Therefore, using data augmentation schemes to mitigate the challenges of small datasets with many categories aligns with the practical needs of urban environmental sound classification tasks.

Table 4. Combined data augmentation ablation experiments.

Model	Technical Method	ESC-50 (%)	UrbanSound8K (%)
AST	-	92.25	88.68
AST	Traditional	94.25	89.48
AST	DRC	94.50	89.73
AST	DRC + TW (10)	95.00	90.32
AST	DRC + TW (20)	95.50	90.92
AST	DRC + TW (30)	94.25	90.20

4.3. Performance Analysis of Transformer Based on Patch-Mix and Contrastive Learning

To verify the superiority of the proposed Patch-Mix and CL-Transformer model framework, a series of ablation experiments are meticulously designed and executed. The models used in these experiments are all pre-trained models and are fine-tuned for 50 epochs. The input features, Aug-Fbank, are enriched through the application of our combined data augmentation technique. Considering that excessive background noise can interfere with the model's learning and recognition of audio signals, while inadequate levels fail to augment the model's noise robustness, we experimentally determined the optimal signal-to-noise ratio (SNR) for the Noise-Fbank feature to be 10 dB. An effective interplay between noise reduction and model robustness is guaranteed by our novel methods.

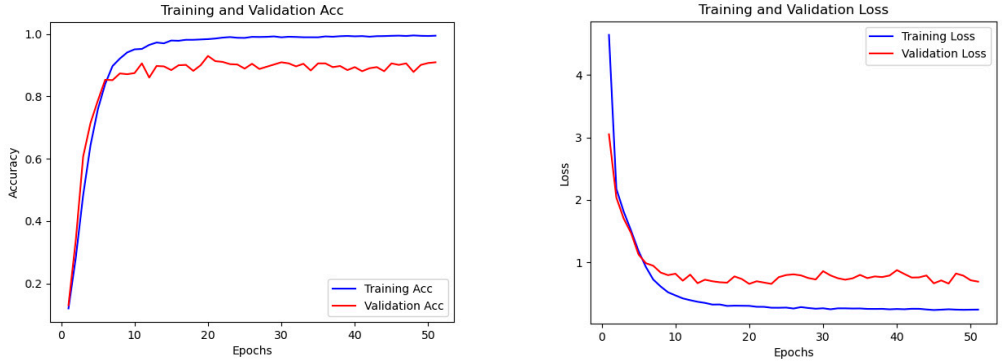
Table 5 showcases the profound impact of incorporating Patch-Mix, utilizing a signal-to-noise ratio (SNR) of 10 dB, into the Fbank feature representation, when coupled with our innovative combined data augmentation technique. This synergy yields outstanding accuracies of 96.5% and 91.63% across the respective datasets. It makes a noteworthy improvement of 1% and 0.71%, respectively. It is supposed to be tailored to the unique requirements of diverse application scenarios. This underscores the effectiveness of our approach in enhancing classification performance in real-world situations. When using the CL-Transformer model as the classification model and applying combined data augmentation to the original audio, the classification accuracies reach 97% and 91.87%, respectively, representing an improvement of 4.75% and 3.19% compared to the baseline model. Finally, by employing the proposed audio classification framework that combines the CL-Transformer model with Patch-Mix, along with the improved combined data augmentation scheme, classification accuracies of 97.75% and 92.95% are achieved on the ESC-50 and UrbanSound8K datasets, respectively. These achievements signify an astonishing improvement of 5.5% and 4.27% over the baseline model, reaching the highest classification accuracy. These experiments demonstrate that Patch-Mix, contrastive learning, and the improved combined data augmentation scheme all contribute to the improvement of classification accuracy.

Figure 5 shows the classification accuracy and loss values on the training and validation sets at different epochs for the UrbanSound8K dataset. It can be clearly observed that the model reaches optimal performance after the 20th epoch. Additionally, the accuracy on both the training and validation sets becomes stable and close to 1, indicating that the proposed method has good generalization ability. Moreover, the proposed model framework

effectively mitigates the overfitting problem commonly encountered by pre-trained models in downstream tasks.

Table 5. Ablation experiments based on Patch-Mix and contrastive learning.

Architecture	Loss	Input Feature	SNR (dB)	Mix	ESC-50 (%)	UrbanSound (%)
AST	CE	Fbank	-	-	92.25	88.68
AST	CE	Aug-Fbank	-	-	95.50	90.92
AST	CE	Aug-Fbank, Fbank	-	Patch	94.25	90.32
AST	CE	Aug-Fbank, Noise-Fbank	0	Patch	94.00	89.61
AST	CE	Aug-Fbank, Noise-Fbank	10	Patch	96.50	91.63
CL-Transformer	CE + CL	Aug-Fbank	-	-	97.00	91.87
CL-Transformer	CE + CL	Aug-Fbank, Noise-Fbank	10	Patch	97.75	92.95



(**a**) (**b**)

Figure 5. Evaluation of accuracy and loss on training data and validation data. (**a**) Accuracy evaluation; (**b**) Loss Evaluation.

4.4. The Impact of the Classification Model Framework on Model Performance

The classification model framework presented in this manuscript ingeniously incorporates Patch-Mix as the feature fusion module, the CL-Transformer model as the cornerstone for classification module. They are vividly combined with each other according to our novel combined data augmentation strategy. To verify the performance of this classification model framework, a comparative analysis is conducted against several state-of-the-art models, utilizing the publicly available ESC-50 and UrbanSound8K datasets as benchmarks. The outcomes of these experiment results, summarized in Table 6 below, provide key evidence of our method's superiority.

Table 6. Comparative experiments of different classification model frameworks.

Classification Model	Method	ESC-50 (%)	UrbanSound8K (%)
DenseNet [22]	DenseNet-201	91.25	85.31
SoundCLR [12]	ResNet + Hybrid Loss	93.40	86.49
AST [20]	Pre-trained Transformer	95.60	-
Transformers [15]	AST + Adamax	95.80	89.80
EAT [36]	End-to-end	96.30	90.00
HTS-AT [37]	Hierarchical Transformer	97.00	-
CL-Transformer	Patch-Mix + Sup-cont	97.75	92.95

The baseline model AST, after fine-tuning, achieved classification accuracies of 92.25% and 88.68% on the ESC-50 and UrbanSound8K datasets, respectively. The CL-Transformer

classification model framework proposed in this paper have achieved classification accuracies of 97.75% and 92.95% on the ESC-50 and UrbanSound8K datasets, respectively, representing improvements of 5.5% and 4.27% over the baseline model and outperforming some current mainstream classification schemes. Figures 6 and 7 vividly illustrate the confusion matrices of our novel CL-Transformer model applied to the ESC-50 and UrbanSound8K sets, respectively. In the figures, the horizontal and vertical coordinates correspond to the audio category labels. From Figure 6, it can be seen that the proposed method achieves 100% accuracy for some typical sound categories in urban environments, such as animal calls, wind noise, thunder, and sirens. From Figure 7, it is evident that the model maintains high classification accuracy across all categories included in the UrbanSound8K dataset.

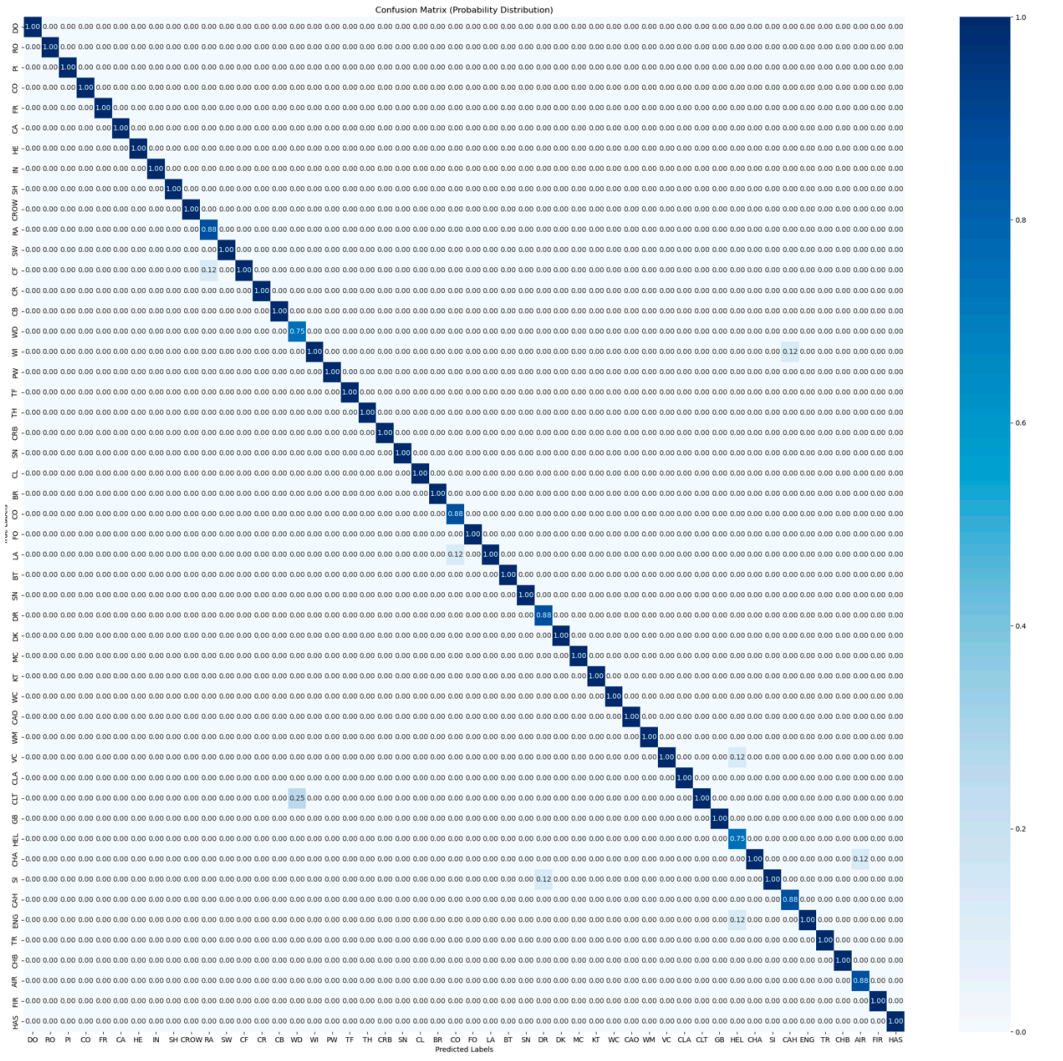


Figure 6. Confusion matrix of the CL-Transformer framework on the ESC-50 validation set. To enhance visualization, we mapped the label values on the horizontal and vertical axes to the abbreviations of each category. DO: Dog; RO: Rooster; PI: Pig; CO: Cow; FR: Frog; CA: Cat; HE: Hen; IN: Insects; SH: Sheep; CROW: Crow; RA: Rain; SW: Sea waves; CF: Crackling fire; CR: Crickets; CB: Chirping birds; WD: Water drops; WI: Wind; PW: Pouring water; TF: Toilet flush; TH: Thunderstorm; CRB: Crying baby; SN: Sneezing; CL: Clapping; BR: Breathing; CO: Coughing; FO: Footsteps; LA: Laughing; BT: Brushing teeth; SN: Snoring; DR: Drinking; DK: Door knock; MC: Mouse click; KT: Keyboard typing; WC: Wood creaks; CAO: Can opening; WM: Washing machine; VC: Vacuum cleaner; CLA: Clock alarm; CLT: Clock tick; GB: Glass breaking; HEL: Helicopter; CHA: Chainsaw; SI: Siren; CAH: Car horn; ENG: Engine; TR: Train; CHB: Church bells; AIR: Airplane; FIR: Fireworks; HAS: Hand saw.

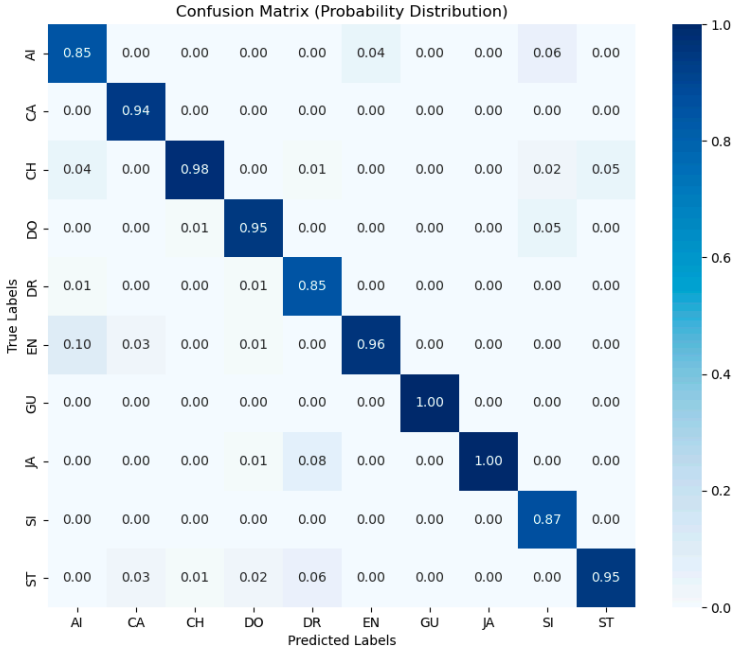
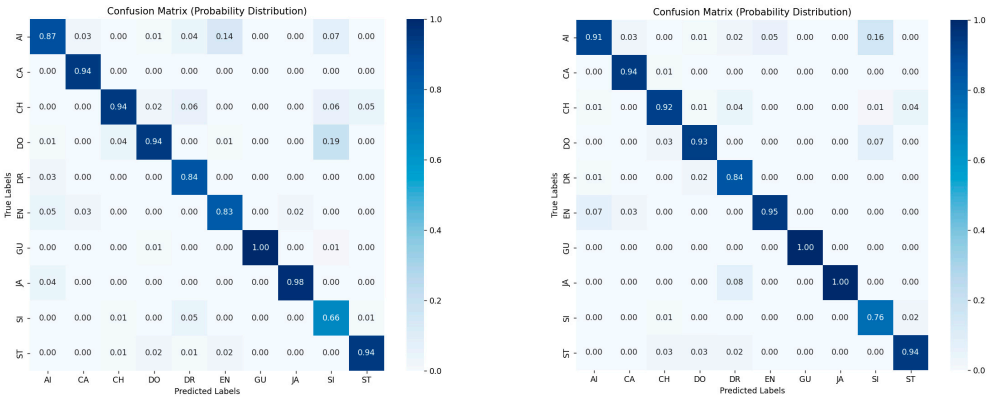


Figure 7. Confusion matrix of the CL-Transformer framework on the UrbanSound8K validation set. AI: air_conditioner; CA: car_horn; CH: children_playing; DO: dog_bark; DR: drilling; EN: engine_idling; GU: gun_shot; JA: jackhammer; SI: siren; ST: street_music.

4.5. Robustness of the CL-Transformer Classification Model Framework to Noise

To assess whether the CL-Transformer model trained with Patch-Mix mixed noise can better handle noise in real-world scenarios, this paper conducts evaluation experiments on the UrbanSound8K dataset. First, we apply noise processing to the UrbanSound8K dataset, adding AWGNs with SNRs of 0 dB and 10 dB to a total of 8732 audio samples. Secondly, we extract the Fbank features of the audio as input for the CL-Transformer model. The experimental results are shown in Figure 8 and Table 7.



(**a**) (**b**)

Figure 8. Confusion matrices on the UrbanSound8K validation set with different SNRs. (**a**) Confusion matrix with SNR of 0 dB; (**b**) Confusion matrix with SNR of 10 dB.

Table 7. Comparison results based on different SNR.

Classification Model	SNR (dB)	Accuracy (%)
CL-Transformer	-	92.95
CL-Transformer	0	88.65
CL-Transformer	10	91.28

The confusion matrices reveal that the innovative audio classification framework attains remarkable classification accuracies of 88.65% and 91.28% on the UrbanSound8K dataset, when subjected to AWGNs with SNRs of 0 dB and 10 dB, respectively. The model's performance surpasses the classification accuracy of some models on the UrbanSound8K dataset without background noise. The experiments demonstrate that our proposed CL-Transformer model framework significantly improves model performance, generalization, and robustness in noisy urban environments. Additionally, it verifies the feasibility of applying this classification framework in real-world scenarios.

5. Conclusions

This paper demonstrates that using the pre-trained AST model and our novel CL-Transformer methods for urban environmental sound classification can obtain better classification accuracy. However, existing methods cannot easily cope with the structural issues of noisy environments. Additionally, pre-trained models have larger parameter sizes, which may predispose them to local minima entrapment during training thread where the model becomes stuck in a relatively poor optimization state, causing its performance to stagnate and resulting in overfitting when classifying sounds. Therefore, this paper proposes an improved combined data augmentation scheme to aid in model training. Furthermore, a novel patch-level feature fusion scheme, Patch-Mix, is introduced to provide more comprehensive feature information to the model. A classification model, CL-Transformer, is also proposed, which introduces supervised contrastive loss to address the overfitting dilemma of pre-trained models and enhance the model's robustness to noise. Extensive experiments conducted on publicly available datasets serve as vivid testament to the robust performance and exceptional noise tolerance ability displayed by our novel methods in the real-world situation.

However, in urban environments, sound signals are more complex. Audio is not only affected by noise interference but also by other sounds such as music, thunder, and wind. In such cases, the capability of the audio classification framework is extremely limited. Therefore, in future work, we will center around the contrastive learning method's basic principles. Our approach will aim to discern the unique representational information of different audio categories within the latent space. So that it is able to reduce the disruptive effects of flexible audio signals and enhance the performance of our audio classification framework. Additionally, transfer learning itself presents some technical challenges, such as how to select an appropriate source domain and how to effectively fine-tune the model. These obstacles may be even more pronounced when applied in urban environments, underscoring the need for innovative strategies to overcome them in the future.

Author Contributions: Conceptualization, X.C. and M.W.; software, X.C. and M.W.; experiment analysis, X.C., M.W. and R.K.; writing—review and editing, X.C., M.W. and H.Q.; funding acquisition, M.W. and H.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This research work is supported by the National Natural Science Foundation of China under Grants 62071135. Guangxi Science and Technology Major Program under Grant No. GuikeAA23062035-2. It is also supported by the Innovation Project of GUET Graduate Education No. 2023YCXB05.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. ESC-50: https://github.com/karolpiczak/ESC-50 (accessed on 25 August 2024). UrbanSound8K: https://urbansounddataset.weebly.com/urbansound8k.html (accessed on 25 August 2024).

Acknowledgments: We are very grateful to volunteers from GLUT for their assistance in the experimental part of this manuscript.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Roozbehi, Z.; Narayanan, A.; Mohaghegh, M.; Saeedinia, S.-A. Dynamic-Structured Reservoir Spiking Neural Network in Sound Localization. *IEEE Access* **2024**, *12*, 24596–24608. [CrossRef]
2. Warnecke, M.; Litovsky, R.Y. Signal Envelope and Speech Intelligibility Differentially Impact Auditory Motion Perception. *Sci. Rep.* **2021**, *11*, 15117.
3. Yang, C.; Gan, X.; Peng, A.; Yuan, X. ResNet Based on Multi-Feature Attention Mechanism for Sound Classification in Noisy Environments. *Sustainability* **2023**, *15*, 10762. [CrossRef]
4. Bansal, A.; Garg, N.K. Environmental Sound Classification: A Descriptive Review of the Literature. *Intell. Syst. Appl.* **2022**, *16*, 200115. [CrossRef]
5. Zhang, D.; Zhong, Z.; Xia, Y.; Wang, Z.; Xiong, W. An Automatic Classification System for Environmental Sound in Smart Cities. *Sensors* **2023**, *23*, 6823. [CrossRef]
6. Mushtaq, Z.; Su, S. Environmental Sound Classification Using a Regularized Deep Convolutional Neural Network with Data Augmentation. *Appl. Acoust.* **2020**, *167*, 107389.
7. Zhang, Z.; Xu, S.; Zhang, S.; Qiao, T.; Cao, S. Attention-Based Convolutional Recurrent Neural Network for Environmental Sound Classification. *Neurocomputing* **2021**, *453*, 896–903. [CrossRef]
8. Al-Hattab, Y.A.; Zaki, H.F.; Shafie, A.A. Rethinking Environmental Sound Classification Using Convolutional Neural Networks: Optimized Parameter Tuning of Single Feature Extraction. *Neural Comput. Appl.* **2021**, *33*, 14495–14506. [CrossRef]
9. Hu, F.; Song, P.; He, R.; Yan, Z.; Yu, Y. MSARN: A Multi-Scale Attention Residual Network for End-to-End Environmental Sound Classification. *Neural Process. Lett.* **2023**, *55*, 11449–11465. [CrossRef]
10. Das, J.K.; Ghosh, A.; Pal, A.K.; Dutta, S.; Chakrabarty, A. Urban Sound Classification Using Convolutional Neural Network and Long Short Term Memory Based on Multiple Features. In Proceedings of the 2020 Fourth International Conference on Intelligent Computing in Data Sciences (ICDS), Hong Kong, China, 27–29 July 2020; pp. 1–9.
11. Mushtaq, Z.; Su, S.F. Efficient Classification of Environmental Sounds through Multiple Features Aggregation and Data Enhancement Techniques for Spectrogram Images. *Symmetry* **2020**, *12*, 1822. [CrossRef]
12. Nasiri, A.; Hu, J. SoundCLR: Contrastive Learning of Representations for Improved Environmental Sound Classification. *arXiv* **2021**, arXiv:2103.01929.
13. Qiao, T.; Zhang, S.; Cao, S.; Xu, S. High Accurate Environmental Sound Classification: Sub-Spectrogram Segmentation versus Temporal-Frequency Attention Mechanism. *Sensors* **2021**, *21*, 5500. [CrossRef] [PubMed]
14. Guzhov, A.; Raue, F.; Hees, J.; Dengel, A. ESResNet: Environmental Sound Classification Based on Visual Domain Models. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Milan, Italy, 10–15 January 2021; pp. 4933–4940.
15. Nogueira, A.F.R.; Oliveira, H.S.; Machado, J.J.M.; Tavares, J.M.R.S. Transformers for Urban Sound Classification—A Comprehensive Performance Evaluation. *Sensors* **2022**, *22*, 8874. [CrossRef] [PubMed]
16. Tsalera, E.; Papadakis, A.; Samarakou, M. Comparison of Pre-Trained CNNs for Audio Classification Using Transfer Learning. *J. Sens. Actuator Netw.* **2021**, *10*, 72. [CrossRef]
17. Kong, Q.; Xu, Y.; Wang, W.; Plumbley, M.D. Sound Event Detection of Weakly Labelled Data with CNN-Transformer and Automatic Threshold Optimization. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2450–2460. [CrossRef]
18. Akbari, H.; Yuan, L.; Qian, R.; Chuang, W.H.; Chang, S.F.; Cui, Y.; Gong, B. VATT: Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text. *Adv. Neural Inf. Process. Syst.* **2021**, *34*, 24206–24221.
19. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st International Conference on Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017.
20. Gong, Y.; Chung, Y.A.; Glass, J. AST: Audio Spectrogram Transformer. *arXiv* **2021**, arXiv:2104.01778.
21. Elliott, D.; Otero, C.E.; Wyatt, S.; Martino, E. Tiny Transformers for Environmental Sound Classification at the Edge. *arXiv* **2021**, arXiv:2103.12157.
22. Salamon, J.; Bello, J.P. Deep Convolutional Neural Networks and Data Augmentation for Environmental Sound Classification. *IEEE Signal Process. Lett.* **2017**, *24*, 279–283. [CrossRef]
23. Park, D.S.; Chan, W. SpecAugment: A New Data Augmentation Method for Automatic Speech Recognition. Google AI Blog, 2019. Available online: https://towardsdatascience.com/state-of-the-art-audio-data-augmentation-with-google-brains-specaugment-and-pytorch-d3d1a3ce291e (accessed on 25 August 2024).
24. Özseven, T. Investigation of the Effectiveness of Time-Frequency Domain Images and Acoustic Features in Urban Sound Classification. *Appl. Acoust.* **2023**, *211*, 109564. [CrossRef]
25. Ashurov, A.; Zhou, Y.; Shi, L. Environmental Sound Classification Based on Transfer-Learning Techniques with Multiple Optimizers. *Electronics* **2022**, *11*, 2279. [CrossRef]
26. Ashurov, A.; Yi, Z.; Liu, H.; Yu, Z.; Li, M. Concatenation-Based Pre-Trained Convolutional Neural Networks Using Attention Mechanism for Environmental Sound Classification. *Appl. Acoust.* **2024**, *216*, 109759. [CrossRef]
27. Jahangir, R.; Nauman, M.A.; Alroobaea, R.; Almotiri, J.; Malik, M.M.; Alzahrani, S.M. Deep Learning-Based Environmental Sound Classification Using Feature Fusion and Data Enhancement. *Comput. Mater. Contin.* **2023**, *75*, 1069–1091. [CrossRef]

28. Mateo, C.; Talavera, J.A. Bridging the Gap between the Short-Time Fourier Transform (STFT), Wavelets, the Constant-Q Transform and Multi-Resolution STFT. *Signal Image Video Process.* **2020**, *14*, 1535–1543. [CrossRef]
29. Bae, S.; Kim, J.W.; Cho, W.Y.; Baek, H.; Son, S.; Lee, B.; Ha, C.; Tae, K.; Kim, S.; Yun, S.-Y. Patch-Mix Contrastive Learning with Audio Spectrogram Transformer on Respiratory Sound Classification. *arXiv* **2023**, arXiv:2305.14032.
30. Madhu, A.; K, S. EnvGAN: A GAN-Based Augmentation to Improve Environmental Sound Classification. *Artif. Intell. Rev.* **2022**, *55*, 6301–6320.
31. Bahmei, B.; Birmingham, E.; Arzanpour, S. CNN-RNN and Data Augmentation Using Deep Convolutional Generative Adversarial Network for Environmental Sound Classification. *IEEE Signal Process. Lett.* **2022**, *29*, 682–686. [CrossRef]
32. Nogueira, A.F.R.; Oliveira, H.S.; Machado, J.J.M.; Tavares, J.M.R.S. Sound Classification and Processing of Urban Environments: A Systematic Literature Review. *Sensors* **2022**, *22*, 8608. [CrossRef]
33. Piczak, K.J. ESC: Dataset for Environmental Sound Classification. In Proceedings of the 23rd ACM International Conference on Multimedia, Brisbane, Australia, 26–30 October 2015; Association for Computing Machinery: New York, NY, USA, 2015; pp. 1015–1018.
34. Salamon, J.; Jacoby, C.; Bello, J.P. A Dataset and Taxonomy for Urban Sound Research. In Proceedings of the 22nd ACM International Conference on Multimedia, New York, NY, USA, 3–7 November 2014; pp. 1041–1044.
35. Deng, J.; Dong, W.; Socher, R.; Li, L.J.; Li, K.; Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, USA, 20–25 June 2009; pp. 248–255.
36. Gazneli, A.; Zimerman, G.; Ridnik, T.; Sharir, G.; Noy, A. End-to-End Audio Strikes Back: Boosting Augmentations towards an Efficient Audio Classification Network. *arXiv* **2022**, arXiv:2204.11479.
37. Chen, K.; Du, X.; Zhu, B.; Ma, Z.; Berg-Kirkpatrick, T.; Dubnov, S. HTS-AT: A Hierarchical Token-Semantic Audio Transformer for Sound Classification and Detection. In Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2022), Singapore, 23–27 May 2022; pp. 646–650.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.