electronics

MDPI

Article

Improving Audio Classification Method by Combining Self-Supervision with Knowledge Distillation

Xuchao Gong [1], Hongjie Duan [1], Yaozhong Yang [1], Lizhuang Tan [2,3,*], Jian Wang [4] and Athanasios V. Vasilakos [5,*]

[1] Artificial Intelligence Research Institute, Shengli Petroleum Management Bureau, Dongying 257000, China; gongxch.slyt@sinopec.com (X.G.)
[2] Key Laboratory of Computing Power Network and Information Security, Ministry of Education, Shandong Computer Science Center (National Supercomputer Center in Jinan), Qilu University of Technology (Shandong Academy of Sciences), Jinan 250013, China
[3] Shandong Provincial Key Laboratory of Computer Networks, Shandong Fundamental Research Center for Computer Science, Jinan 250013, China
[4] College of Science, China University of Petroleum (East China), Qingdao 266580, China; wangjiannl@upc.edu.cn
[5] Department of ICT, Center for AI Research (CAIR), University of Agder (UiA), 4879 Grimstad, Norway
* Correspondence: tanlzh@sdas.org (L.T.); thanos.vasilakos@uia.no (A.V.V.)

Abstract: The current audio single-mode self-supervised classification mainly adopts a strategy based on audio spectrum reconstruction. Overall, its self-supervised approach is relatively single and cannot fully mine key semantic information in the time and frequency domains. In this regard, this article proposes a self-supervised method combined with knowledge distillation to further improve the performance of audio classification tasks. Firstly, considering the particularity of the two-dimensional audio spectrum, both self-supervised strategy construction is carried out in a single dimension in the time and frequency domains, and self-supervised construction is carried out in the joint dimension of time and frequency. Effectively learn audio spectrum details and key discriminative information through information reconstruction, comparative learning, and other methods. Secondly, in terms of feature self-supervision, two learning strategies for teacher-student models are constructed, which are internal to the model and based on knowledge distillation. Fitting the teacher's model feature expression ability, further enhances the generalization of audio classification. Comparative experiments were conducted using the AudioSet dataset, ESC50 dataset, and VGGSound dataset. The results showed that the algorithm proposed in this paper has a 0.5% to 1.3% improvement in recognition accuracy compared to the optimal method based on audio single mode.

Keywords: audio classification; comparative learning; knowledge distillation; masked auto-encoder; self-supervision; transformer

check for updates

Citation: Gong, X.; Duan, H.; Yang, Y.; Tan, L.; Wang, J.; Vasilakos, A.V. Improving Audio Classification Method by Combining Self-Supervision with Knowledge Distillation. *Electronics* **2024**, *13*, 52. https://doi.org/10.3390/electronics13010052

Academic Editor: Alberto Fernandez Hilario

Received: 26 November 2023
Revised: 16 December 2023
Accepted: 18 December 2023
Published: 21 December 2023

1. Introduction

In recent years, with the rapid development of mobile multimedia technology, the exponential growth of audiovisual data have increased the demand for the capability of audio classification [1]. Audio classification, incorporating relevant technologies from machine learning and signal processing, plays a crucial role in applications such as speech recognition, sound event detection, emotion analysis, music classification, and speaker recognition. The objective of audio classification is to accurately categorize audio signals into predefined classes, facilitating the identification and understanding of different sound sources for improved downstream applications.

Although supervised audio classification methods have demonstrated effectiveness in many scenarios, they heavily rely on extensive labeled data, leading to increased costs in practice. Simultaneously, in numerous experiments, it has been observed that directly applying label-based discrete learning to audio information processing can result in classification bias [2]. The reason is that although the audio signal is continuous, the duration of the same sound event varies in different situations. Uniform discretization learning can easily lead to learning bias. In other words, supervised discrete learning fails to effectively elicit advanced semantic features from continuous audio and discard redundant details.

The reconstruction aspect can be divided into two major categories: spectrogram reconstruction and feature reconstruction. For the former, considering the two-dimensional specificity of audio spectrograms, self-supervised strategies can be constructed in a single dimension of time or frequency or jointly in the time-frequency dimension. For the latter, model-internal or teacher-student model learning strategies based on knowledge distillation can be constructed. By conducting more comprehensive self-supervised learning from these two dimensions, the recognition ability of audio single-modal classification is further optimized.

In light of the comprehensive analysis above, this paper introduces an innovative algorithm for audio classification, termed ACM-SSKD (Audio Classification Method based on Self-Supervised and Knowledge Distillation). The proposed method offers the following advantages:

(1) Multifaceted Self-Supervised Learning Mechanisms: Introducing various self-supervised learning mechanisms based on spectrograms in audio classification, we construct two self-supervised strategies: time-frequency random masking and spectrogram block random masking. Through contrastive learning, we achieve discriminative feature learning, and by self-learning information reconstruction through masking, we effectively capture intricate details in the audio spectrum.

(2) Feature Reconstruction Strategies: In the realm of feature reconstruction, two teacher-student learning strategies are devised. Leveraging knowledge distillation learning mechanisms, these strategies enhance model feature representation, leading to rapid convergence and efficient learning.

(3) Experimental Validation: Experimental results demonstrate the effectiveness of combining spectrogram-based self-supervised strategies for learning intricate audio features. Furthermore, feature reconstruction enhances model learning, yielding excellent results in multiple publicly available audio classification test sets. Notably, in pure audio recognition on the AudioSet-2M, ESC-50, and VGG Sound datasets, the proposed method achieves accuracy rates of 49.9%, 98.7%, and 61.3%, respectively. These results surpass the current state-of-the-art single-modal methods by 1.3%, 0.6%, and 0.5%, respectively.

2. Related Work

Over the past few years, supervised audio classification methods have demonstrated excellent performance in various publicly available datasets [1,3–7]. In the specific modeling process, supervised learning for audio classification assigns a discrete label or category to a segment of continuous audio information. The audio information is then projected through the model to generate feature vectors with rich audio semantics, which are subsequently mapped to discrete labels, ultimately achieving the goal of classification. AST (Audio Spectrogram Transformer) [5], utilizing audio spectrograms as input, employs two-dimensional convolution to extract serialized features, followed by cascaded operations of multiple transformer blocks to obtain global features of the audio sequence and improve recognition performance significantly. Panns [1] leverages the large-scale audio dataset AudioSet for training, exploring various aspects of audio classification effects, such as depth, feature dimensions, dropout ratios, and spectrum generation methods, proposing high-quality models. Considering that CNN (Convolutional Neural Network) [8,9] focuses on the local context, PSLA (Pretraining, Sampling, Labeling, and Aggregation) [10] introduces a pooling attention mechanism for feature enhancement to capture global audio information and improve classification performance. To effectively enhance supervised audio classification, Khaled Koutini et al. [11] decomposes audio Transformer position encoding into temporal and frequency components, supporting variable-length audio classification. Arsha Nagrani [12], combining Transformer modeling, promotes the model's learning ability through an intermediate feature fusion approach. Ke Chen [13] introduces a hierarchical audio Transformer with a semantic module that combines with input tokens, mapping the final output to class feature maps and enhancing classification performance. Eranns [14] reduces computational complexity by introducing model scale hyperparameters. By utilizing optimal parameter values, computational efficiency is improved, leading to potential savings or performance enhancement.

Additionally, to fully leverage the benefits of supervised pre-training and further enhance the effectiveness of audio classification, many methods employ model weights from the image domain. AST [5] initializes its model based on the pre-trained model on ImageNet using VIT [15], effectively boosting model performance. Hts-at [13] utilizes pre-training weights from the Swin Transformer [16] on image datasets, significantly improving audio classification results. PaSST [11] incorporates pre-training weights from DeiT [15]. To effectively confirm the appropriate number of hidden layer nodes in the neural network, Xuetao Xie [17] using the L1 regularization method proposes several effective methods to determine the optimal number of hidden nodes for a perceptron network.

From another perspective, advanced semantic abstraction of continuous audio can be achieved through signal autoencoding-decoding reconstruction. This technique is a powerful means of self-supervised learning. Compared to supervised learning, self-supervised learning does not require a large number of labeled samples. From this perspective, self-supervised data are easily obtainable. Similar to other self-supervised learning methods, self-supervised audio training typically aims to learn its representations through contrastive learning or reconstruction.

Self-supervised techniques have found numerous applications in recent years in audio classification [18–23]. Concurrently, various studies [24–29] indicate that reconstruction-based self-supervised techniques are not only effective for speech but also exhibit robust learning capabilities in modeling information such as video images and multimodal fusion.

Considering that audio often encompasses a variety of environmental events, such as speech, ambient sounds, and musical beats, often accompanied by considerable ambient noise, this poses significant challenges for universal audio classification modeling. In response, approaches like Dading Chong et al. [18] and Hu Xu et al. [19] apply masking operations to spectrograms, in the pre-training stage, and self-supervised acoustic feature reconstruction is used as the pre-training target. COLA [30] to achieve good pre-training results, during the pre-training period, comparative learning is performed on the audio dataset, assigning high similarity to data from the same audio segment, and low similarity to data from different segments. Eduardo Fonseca et al. [31] enhance sound event learning through different view-enhanced learning tasks, demonstrating that unsupervised contrastive pre-training can alleviate the impact of data scarcity and improve generalization. For more effective contrastive learning, Clar [32] proposes several efficient data augmentation and enhancement methods. Luyu Wang [33] introduces a contrastive learning method with audio samples in different formats, maximizing consistency between original audio and its acoustic features. To enhance generalization and obtain a robust audio representation, Daisuke Niizumi [34] trained and learned different audio samples using mean square error loss and exponential moving average optimization strategies. A patch-based self-supervised learning method is proposed by Ssast [24] for pre-training and achieving good performance. Mae-ast [35], based on the Transformer encoding-decoding structure, reconstructs pre-training tasks where the decoder is used for masked reconstruction, demonstrating excellent recognition performance in multi-audio classification tasks. Andrew N Carr [36] sequentially shuffles input audio features, implementing end-to-end model pre-training with a differentiable sorting strategy, and exploring self-supervised audio pre-training methods with masked discrete label prediction targets. In order to effectively distinguish unsupervised features, AARC [37] integrates the selection of unsupervised features and the determination of network structure into a unified framework, while adding two Group Lasso losses to the objective function.

Although self-supervised spectrogram pre-training strategies have shown good results in audio classification, some methods argue that this self-supervised reconstruction is relatively singular and can only restore low-level time-frequency features, with weaker capabilities in advanced audio semantic abstraction [38,39].

3. Multi-Dimensional Self-Supervised Learning

Figure 1 illustrates the training process for audio recognition in this paper, emphasizing two key components: Multi-Dimensional Self-Supervised Learning (Multi-SSL) and Knowledge Distillation.

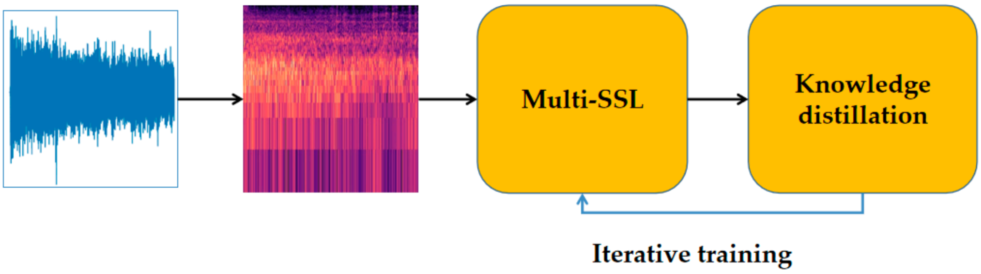

Figure 1. Workflow of the Proposed Method ACM-SSKD.

Initially, one-dimensional audio signals are transformed into two-dimensional Mel spectrograms. During training, in the first iteration, audio samples undergo Multi-SSL operations to acquire detailed internal descriptions through information reconstruction. This forms the basis for training the discrete-label audio classification task, enhancing training efficiency and model generalization through knowledge distillation. The model from the first iteration is used as the initialization, undergoing Multi-SSL training once again to create a new self-supervised model with audio knowledge extraction capabilities. This iterative process is repeated through knowledge distillation training, demonstrating its effectiveness in improving classification outcomes. The detailed algorithm process is shown in Algorithm A1 in Appendix A.

Figure 2 outlines the modeling process for the proposed Multi-Dimensional Self-Supervised Learning model, comprising three main parts: self-supervised information construction, self-supervised modeling, and information reconstruction with feature fitting.

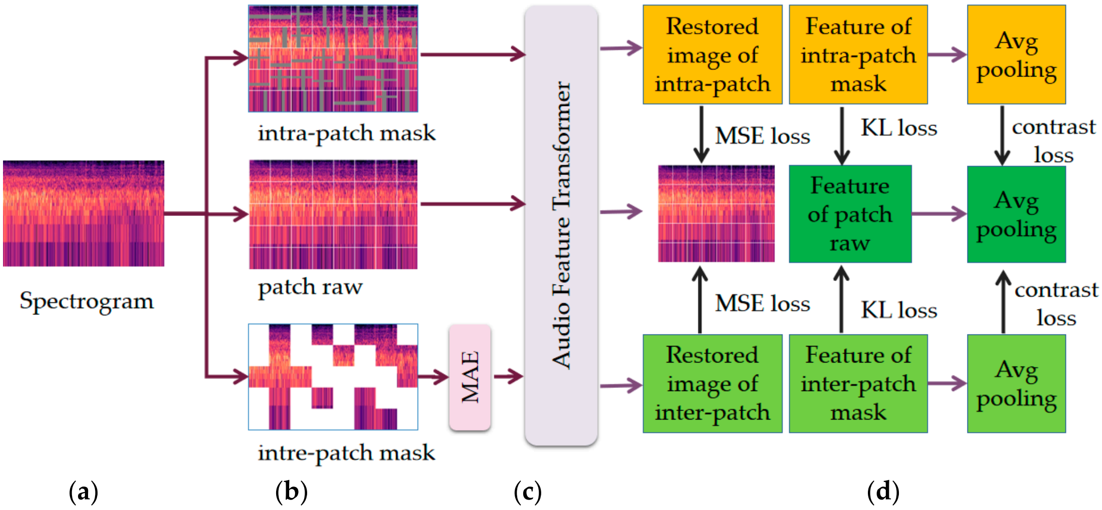

Figure 2. Schematic Diagram of Multidimensional Self-Supervision Model. (**a**) Two-dimensional spectrogram; (**b**) Information Reconstruction and Feature Fitting; (**c**) Multidimensional Self-Supervised Training; (**d**) Information Reconstruction and Feature Fitting.

3.1. Self-Supervised Information Construction

The conversion of one-dimensional audio information into two-dimensional spectrograms facilitates its treatment as an image. Simultaneously, owing to the temporal-frequency transformation during spectrogram conversion, the horizontal axis of the spectrogram can be interpreted as the time dimension, while the vertical axis represents the frequency dimension. Building upon these considerations and drawing inspiration from AST [5], this paper introduces a framework for Multi-Self-Supervised Learning (Multi-SSL). Through the establishment of a time-domain-frequency-domain masked self-supervision mechanism (as depicted in Figure 3), the model learns the inter-domain relationships between different features. Concurrently, with the spectrogram block masked self-supervision technique (illustrated in Figure 4), the aim is to harness minimal domain knowledge to achieve robust representational capabilities—a strategy validated in image self-supervised modeling [40].

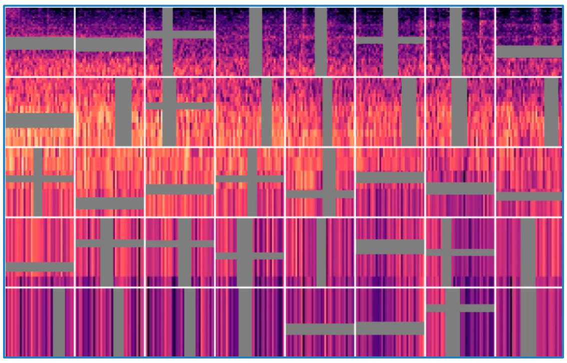

Figure 3. Schematic Diagram of Time-Frequency Masks.

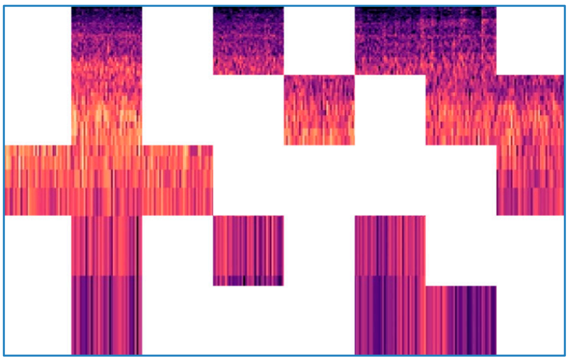

Figure 4. Schematic Diagram of Spectrogram Block Masks.

In the construction of the time-domain-frequency-domain mask (intra-patch mask), technical details from audio processing are incorporated [41]. In contrast to random masking between the two domains, a departure is made to align with the spectrogram block corresponding to the transformer. Random masking is implemented within each patch, as illustrated in Figure 3.

During the spectrogram block masking (inter-patch mask) phase, the entire spectrogram is treated as a whole, enabling reconstruction learning through masking of the entire patch (Figure 4). This approach, proven effective in image recognition [40] and multimodal modeling [25], takes into account the impact of the underlying distribution of low-level image features on mask selection. As depicted in Figure 5, this paper randomly samples a hundred instances from the Audio Set dataset and calculates the distribution of Histogram of Oriented Gradients (HOG) features within each patch. Notably, key audio events often concentrate in positions with high feature distribution variance. Consequently, considering this variance during random sampling proves meaningful, a notion validated in subsequent ablation experiments, as expressed in Formula (1).

$$S_{patch\_mask} = 1/e^{\frac{(x-\mu)^2}{\sigma^2}} \tag{1}$$

Drawing on the mean and variance of HOG feature values within each patch, we construct sampling probabilities. The hypothesis is that events representing sound should exhibit a Gaussian distribution within a certain time step. Accordingly, the closer a sample is to the mean, the higher its sampling probability. In Formula (1), x denotes the gradient mean within each patch, μ signifies the mean gradient across all patches, and σ represents the gradient variance across all patches.

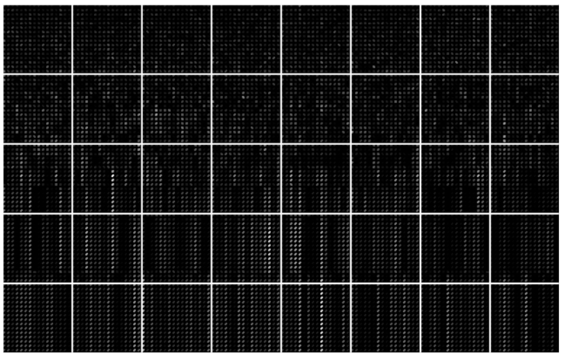

Figure 5. Statistical Analysis of Audio Spectrogram HOG Features.

3.2. Multi-Dimensional Self-Supervised Modeling

During the inter-patch mask phase, as the remaining image blocks decrease in size after random masking, an additional encoding module is introduced. In this regard, the paper draws inspiration from the strategy employed in Image MAE (Masked AutoEncoders) [40]. This modeling process yields the encoding features of the spectrogram blocks (inter-patch feature).

Within the self-supervised modeling framework, the time-frequency-domain mask (intra-patch mask), original information (block segmentation without masking), and spectrogram block encoding features (inter-patch feature) are simultaneously input. Through the Audio Feature Transformer, three serialized features representing the audio are obtained—the feature of the intra-patch mask, the feature of the inter-patch mask, and the feature of the patch raw. Simultaneously, a series of perceptron layers decode the features, resulting in corresponding reconstructed spectrograms (restored image of intra-patch, restored image of raw, restored image of inter-patch).

3.3. Information Reconstruction and Feature Fitting

The subsequent step involves constructing self-supervised losses, encompassing spectrogram reconstruction and feature reconstruction. For the reconstructed spectrograms (restored image of intra-patch, restored image of inter-patch), mean squared error reconstruction (MSE loss) is employed. Similar to the construction of the spectrogram block mask (inter-patch mask), The HOG gradient prior is considered during loss calculation.

$$L_{restore}^{intra-patch\_mask} = \frac{1}{M}\sum_{m=1}^{M}\sum_{w=1}^{W}\sum_{h=1}^{H}\left(\frac{1}{e^{(x-\mu)^2/\sigma^2}}\left(x_{pre} - x_{gt}\right)^2\right) \tag{2}$$

$L_{restore}^{intra-patch\_mask}$ denotes the intra-patch spectrogram reconstruction loss, where x_{pre} signifies the predicted pixel values, and x_{gt} represents the true values of the spectrogram pixels. Similarly, the inter-patch spectrogram reconstruction follows the same approach and can be represented as $L_{restore}^{inter-patch\_mask}$. The overall reconstruction loss is as in Formula (3).

$$L_{restore}^{all} = L_{restore}^{intra-patch\_mask} + L_{restore}^{inter-patch\_mask} \tag{3}$$

In terms of feature reconstruction, given that transformer serialized modeling involves multiple sets of features and includes a single sequence feature for final classification (utilizing AVG pooling in this paper), reconstructed serialized features (feature of intra-patch mask, feature of inter-patch mask, feature of patch raw) are fitted using a combination of mean squared error and KL divergence.

$$L_{mse}^{intra-patch\_mask} = \frac{1}{M}\sum_{m=1}^{M}\sum_{w=1}^{W}\sum_{h=1}^{H}\left(\left(f_{pre} - f_{raw}\right)^{2}\right) \tag{4}$$

$$L_{KL}^{intra-patch\_mask} = \frac{1}{M}\sum_{m=1}^{M}\sum_{w=1}^{W}\sum_{h=1}^{H}\left(\varphi(f_{raw})\frac{\varphi(f_{raw})}{\varphi(f_{pre})}\right) \tag{5}$$

$L_{mse}^{intra-patch\_mask}$ illustrates the use of MSE loss for intra-patch feature reconstruction to fit the overall feature. To bring each dimension of intra-patch features closer to the true values in the distribution, $L_{KL}^{intra-patch\_mask}$ KL loss is introduced, as outlined in Formula (5). f_{pre} denotes the features obtained after audio feature transformer operation for intra-patch, while f_{raw} represents the features acquired from patch raw. It is noteworthy that the features obtained from the image block without masking are treated as the fitting target, aiming to further restore and reconstruct at the feature level. A similar approach is adopted for inter-patch feature reconstruction, with MSE loss represented as $L_{mse}^{inter-patch\_mask}$, and KL loss represented as $L_{KL}^{inter-patch\_mask}$. The overall feature minimization reconstruction loss is expressed in Formula (6).

$$L_{mse}^{all} = L_{mse}^{intra-patch\_mask} + L_{mse}^{inter-patch\_mask} \tag{6}$$

Formula (7) represents the distribution minimization reconstruction loss for features.

$$L_{KL}^{all} = L_{KL}^{intra-patch\_mask} + L_{KL}^{inter-patch\_mask} \tag{7}$$

The final feature for audio classification is a single sequence feature obtained through AVG pooling. To enhance discriminability, contrastive learning (Contrast loss) is employed.

$$L_{contrast}(pre, y) = \frac{1}{B}\sum_{b=1}^{B} log\frac{exp(C(pre_{b}, y_{b})/\tau)}{\sum\limits_{k=1}^{B} exp(C(pre_{b}, y_{k})/\tau)} \tag{8}$$

Formula (8) outlines the contrastive learning strategy adopted in this paper. Here, $C(pre_b, y_k) = pre_b^T y_k / \|pre_b^T\| \|y_k\|$ indicates comparison through cosine similarity between the predicted feature pre_b and the fitted feature y_k, τ denotes the temperature control parameter, B represents the batch size during training, and $L_{contrast}$ denotes the modeled contrast loss.

In specific experiments, intra-patch and inter-patch are considered as data augmentations for patch raw. The constructed contrastive loss function is expressed in Formula (9).

$$L_{contrast}^{all} = L_{contrast}\left(pre_{intra-patch}, pre_{patch-raw}\right) \quad + L_{contrast}\left(pre_{inter-patch}, pre_{patch-raw}\right) \tag{9}$$

In summary, this paper’s self-supervised approach involves two functionalities (spectrogram reconstruction and feature fitting) and three types of loss functions (MSE loss, KL loss, and Contrast loss). Considering the task’s relevance in specific experiments, similar learning tasks are assigned equal weights. Thus, the eight loss functions can be consolidated into three groups, as outlined in Formula (10).

$$L_{total} = \alpha\left(L_{restore}^{all} + L_{mse}^{all}\right) + \beta L_{KL}^{all} + L_{contrast}^{all} \tag{10}$$

In the above equation, α and β, respectively represent the weights assigned to each group of loss functions.

3.4. Knowledge Distillation

As shown in Figure 6, knowledge distillation based on the teacher model, can be regarded as an expedient means to accelerate iterations, leveraging the exemplary BEATS model [21] as the target for feature fitting in the context of supervised learning with discrete labels.

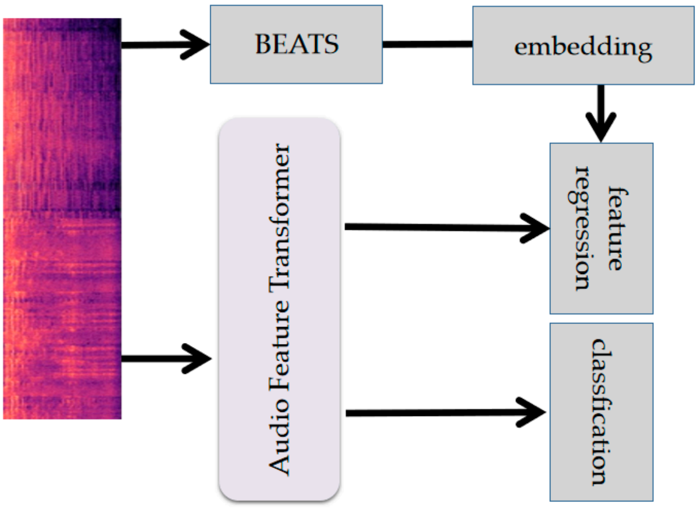

Figure 6. Supervised Learning Combined with Knowledge Distillation.

During training, to guide the network towards a more stable convergence, the supervised cross-entropy classification loss and the self-supervised feature fitting loss are combined.

$$L_{cls} = L_{CE}(F_{cls}, F_{teacher}) \tag{11}$$

$$L_{regress} = L_{mse}(F_{cls}, F_{teacher}) \tag{12}$$

$$L = L_{cls} + \gamma L_{regress} \tag{13}$$

Here, L_{cls} represents the classification cross-entropy loss, $L_{regress}$ is the feature regression loss, F_{cls} denotes the features obtained through the audio feature transformer network mentioned in Section 3.2, and Fteacher signifies the features being fitted, obtained through BEATS. The parameter γ serves as the loss weight control factor.

4. Experiments

To assess the effectiveness of the proposed method in this paper, three datasets Audioset [42], VGGSound [27], and ESC-50 [43], were employed for evaluation. AudioSet is a large-scale audio classification dataset comprising over two million 10-s YouTube clips, spanning 527 audio categories. Each sample contains one or more audio categories, and the dataset is partitioned into class-balanced sets (22 K samples), class-imbalanced sets (2000 K samples), and an evaluation set (20 K samples). To address the temporal nature of YouTube videos, this study downloaded and parsed 20 K class-balanced sets, 1900 K class-imbalanced sets, and 18 K evaluation sets, aligning with prior work [21]. VGGSound encompasses 200 K 10-s audio-video clips with 309 audio categories, featuring 183 K training samples and 15 K testing samples. ESC-50, an environmental sound classification dataset, includes 2000 5-s audio samples across 50 audio categories.

4.1. Impact of Multi-Dimensional Self-Supervised Tasks on Classification

To further elucidate the universality of incorporating self-supervised tasks in classification recognition, this paper conducts experiments on both the intra-patch self-supervision strategy (intra-SS) itself and the corresponding improvements and combinations of loss functions. Similarly, experiments are conducted on the inter-patch self-supervision strategy (inter-SS) itself and the corresponding combinations of loss functions.

In the experiments on the intra-patch feature map self-supervision strategy (intra-SSR), concerning all sequence blocks of the spectrogram, this paper sets the number of spectrogram blocks masked in each iteration to not exceed 50% of the total patch count. Within each masked spectrogram block, there is a 30% probability of temporal masking, a 30% probability of frequency masking, and a 40% probability of joint temporal-frequency masking. The range of frequency-domain and temporal-domain masking does not exceed 25% of each spectrogram block.

The notation "Audioset 20K" denotes that, during the discrete label training phase, 20,000 samples are randomly extracted from the Audioset dataset for training, with 10,000 samples used for evaluation. Similarly, "VGGSound 20K" indicates that, during the discrete label training phase, 20,000 samples are randomly selected from the VGGSound training set for training, and 10,000 samples are randomly selected from the test set for evaluation.

In Table 1, "Base" represents the baseline effect without an intra-patch self-supervision strategy. After the incorporation of the self-supervision strategy (intra-SSR), there is a gain of 0.3% and 0.5% in Audioset and VGGSound, respectively. By introducing HOG feature priors in the reconstruction loss function (intra-SSR and HOG weight), further improvements are observed in terms of accuracy.

Table 1. Impact of Intra-Patch Spectrogram Reconstruction on Classification Performance.

Comparison Metrics	Base	+Intra-SSR	+HOG Weight
Audioset 20K	0.371	0.374	0.377
VGGSound 20K	0.537	0.542	0.544

When validating the intra-patch feature reconstruction self-supervision strategy (intra-SSF), the experimental data, input spectrogram settings, and "Base" representing the baseline results are consistent with Table 2. Incorporating the feature fitting MSE loss yields varying degrees of improvement on both datasets. The addition of feature distribution learning KL loss does not manifest improvement conclusions in the Audioset 20K dataset, while a gain of 0.2% is observed in VGGSound 20K, suggesting that the generalizability of feature fitting in audio classification tasks may not be consistent but potentially lacks side effects. The improvement is relatively pronounced after introducing the contrast loss.

Table 2. Impact of Intra-Patch Feature Reconstruction on Classification Performance.

Loss Terms	Audioset 20K	VGGSound 20K
Base	0.377	0.544
+intra-SSF & mse	0.383	0.549
+intra-SSF & KL	0.383	0.551
+intra-SSF & cont	0.391	0.557

In the experiments on the inter-patch feature map self-supervision strategy (inter-SSR), the initial exploration focused on the impact of varying proportions of masking in the MAE encoding module on the ultimate classification performance. Table 3 showcases the influence of different masking proportions within the MAE module on classification results, revealing optimal stability at a masking ratio of 75%. This proportion was subsequently employed in the specific classification experiments.

Table 3. Impact of Inter-Patch Mask Ratios on Classification Performance.

Mask Ratios	45%	55%	65%	75%	85%
Audioset 20K	0.353	0.361	0.363	0.365	0.364
VGGSound 20K	0.526	0.539	0.541	0.542	0.542

Within the context of the inter-patch feature map self-supervision strategy (inter-SSR), experiments were conducted using the Audioset and VGGSound datasets. In order to visually observe the differences in HOG features in two-dimensional spectrograms of different categories, we randomly selected four scenarios: train, machine, ship, and pigeon, as shown in Figure 7, from top to bottom, they represent video images, one-dimensional audio raw data, two-dimensional audio spectrograms, and spectrograms extracted through HOG features. From the fourth line, we can see that after introducing HOG features prior in the spectrogram highlight key regions and enhance the ability to distinguish between different categories.

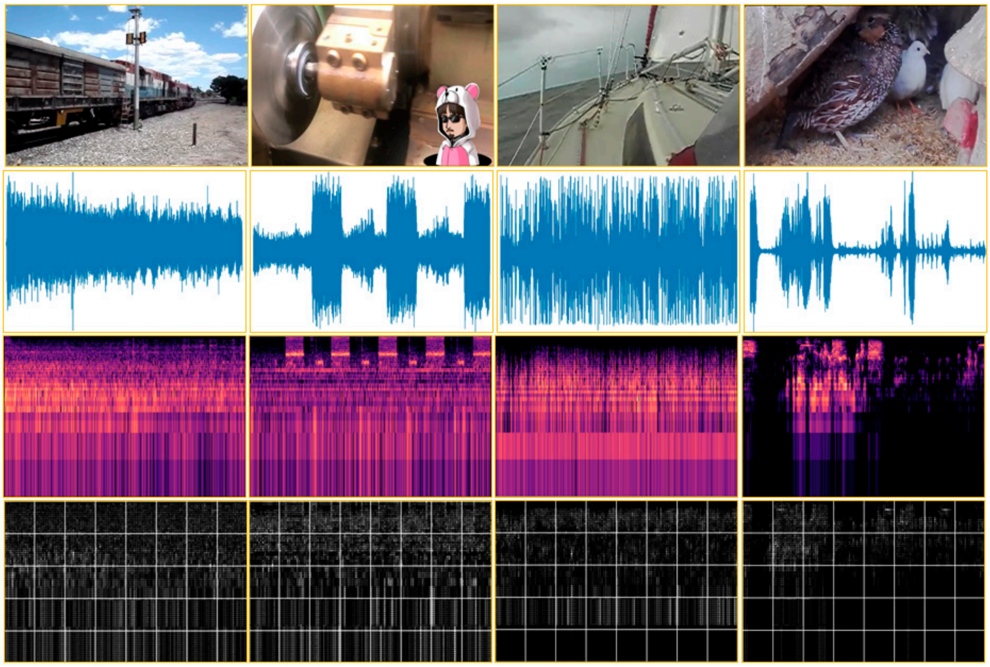
(**a**) Train (**b**) Machine (**c**) Ship (**d**) Pigeon

Figure 7. Differences in Audio Spectrogram HOG Features Across Different Sound Categories.

Table 4 demonstrates the impact of incorporating the spectrogram HOG prior as a weight in the training loss for audio classification. The results indicate further improvements in recognition metrics following the introduction of inter-patch feature map self-supervision and the HOG feature prior.

Table 4. Impact of Inter-Patch Spectrogram Reconstruction on Classification Performance.

Loss Functions	Base	+Inter-SSR	+HOG Weight
Audioset 20K	0.371	0.379	0.382
VGGSound 20K	0.537	0.544	0.546

When validating the inter-patch feature reconstruction self-supervision strategy (inter-SSF), the experimental data and input spectrogram settings remained consistent with Table 5. "Base" represents the baseline results based on inter-patch spectrogram reconstruction. The incorporation of MSE loss for feature fitting resulted in varying degrees of improvement on both datasets. The addition of feature distribution learning KL loss yielded slight gains on both datasets, further emphasizing the role of feature fitting in audio classification tasks. The improvement was more pronounced after introducing the contrast loss.

Table 5. Impact of Inter-Patch Feature Reconstruction on Classification Performance.

Loss Functions	Audioset 20K	VGGSound 20K
Base	0.382	0.546
+inter-SSF & mse	0.387	0.551
+inter-SSF & KL	0.389	0.552
+inter-SSF & cont	0.394	0.561

To further assess the benefits of jointly modeling intra-patch and inter-patch features for audio classification results, this paper conducted corresponding explorations. Results in Table 6 show that the combination of both approaches can further enhance the effectiveness of audio classification.

Table 6. Impact of Multidimensional Self-Supervision on Classification Performance.

Self-Supervision Approaches	Audioset 20K	VGGSound 20K
Base	0.371	0.537
intra-SS	0.391	0.557
inter-SS	0.394	0.561
intra-SS & inter-SS	0.402	0.566

To validate the efficacy of the proposed multi-dimensional self-supervision, experiments in multi-dimensional self-supervised training were conducted based on random initialization for all aforementioned self-supervision experiments. To investigate the potential enhancement in final results by utilizing pre-trained weights from existing models, this paper explored the use of pre-training weights from ViT on ImageNet as initialization parameters in subsequent experiments, as indicated in Table 7. "Random Init" denotes the initial use of random initialization during self-supervised training, while "ImageNet Init" signifies the use of pre-training weights from ViT on ImageNet. Results suggest the effectiveness of employing weights from different domains for audio classification.

Table 7. Influence of Existing Weight Initialization on Multidimensional Self-Supervision.

Initialization Methods	Audioset 20K	VGGSound 20K
Random Init	0.402	0.566
ImageNet Init	0.404	0.570

4.2. *Impact of Audio Knowledge Distillation on Classification*

To further enhance the effectiveness of audio classification, we incorporated knowledge distillation based on state-of-the-art models in addition to training with discrete labels. In our specific experiments, we initialized the process with various parameters, including AST [5], BEATS [21], and SPFA (Self-supervision with Parameter-Free Attention) [44], aiming to validate the improvement in audio classification by employing different teacher models as fitting targets. As shown in Table 8, the inclusion of teacher models resulted in a notable enhancement in recognition performance, with higher-performing teacher models yielding greater improvements in our algorithm. It is essential to note that, for a more objective comparison, and to validate the effectiveness of our algorithm framework, AST [5] was chosen as the teacher model in the knowledge distillation process during the specific experiments.

Table 8. Impact of Different Teacher Models on Classification Performance.

Teacher Model Categories	Audioset 20K	VGGSound 20K
None	0.402	0.566
AST	0.405	0.568
SPFA	0.406	0.571
BEATS	0.409	0.575

As previously described, this paper leveraged multidimensional self-supervision to obtain pre-trained models. The audio classification was carried out through an alternating process involving knowledge distillation and discrete label learning. Consequently, we explored the impact of different iteration counts in these two distinct phases on recognition results. In this experiment, BEATS was employed as the teacher model, and the results in Table 9 indicate that multiple iterations of alternating training contribute to further improvements in outcomes.

Table 9. Effect of Different Iteration Counts on Classification Performance.

Iteration Counts	Audioset 20K	VGGSound 20K
iter1	0.409	0.575
iter2	0.415	0.583
iter3	0.417	0.586

4.3. Comparative Analysis of Classification Results Using Different Methods

When evaluating the recognition performance among various methods, this study conducted experiments on the aforementioned three datasets. Table 10 presents a comparative analysis of the ACM-SSKD method proposed in this paper and various other methods on the Audioset dataset. It is noteworthy that, for a more objective comparison, AST [5] was chosen as the teacher model in the knowledge distillation process, and in the table, iter1–3 represents the number of iterations in the alternating training of multidimensional self-supervision and knowledge distillation. The results indicate that under the same number of iterations, our approach achieves recognition metrics 0.3–0.8% higher than the BEATS method. Ensemble refers to the fusion of results from the three models in the experiment, following a fusion approach consistent with AST.

Table 10. Comparative Performance of Different Methods on the Audioset Dataset.

Method	Model Param	Pre-Trained Data	Audioset
PANN [1]	81M	-	0.431
PSLA [10]	14M	ImageNet	0.444
ERANN [14]	55M	-	0.450
AST [5]	86M	ImageNet + AudioSet	0.459
PaSST [11]	86M	ImageNet + AudioSet	0.471
Hts-at [13]	31M	ImageNet + AudioSet	0.471
MaskedSpec [18]	86M	AudioSet	0.471
CAV-MAE [45]	94M	ImageNet + AudioSet	0.449
SPFA (Single) [44]	87M	-	0.464
BEATS (iter1) [21]	90M	AudioSet	0.479
BEATS (iter2) [21]	90M	AudioSet	0.481
BEATS (iter3) [21]	90M	AudioSet	0.480
BEATS (iter3+) [21]	90M	AudioSet	0.486
ACM-SSKD (iter1)	92M	ImageNet + AudioSet	0.483
ACM-SSKD (iter2)	92M	ImageNet + AudioSet	0.486
ACM-SSKD (iter3)	92M	ImageNet + AudioSet	0.492
ACM-SSKD (Ensemble)	92M	ImageNet + AudioSet	0.499

The Model Parameters column in Table 10 represents the number of modeling parameters, and it can be seen that the parameter quantity in this paper is slightly higher than BEATS and lower than CAV-MAE. The Pre-training Data column represents the dataset used during pre-training. Combined with the weight initialization experiment in Table 7 of this paper, it can be inferred to some extent that pre-trained with image modality data can also promote the improvement of classification performance, similar conclusions can also be drawn from Tables 11 and 12. From the perspective of correlation, imagenet does not have a strong correlation with the current audio dataset, but pre-training based on it can still improve the effectiveness. If combined with the video content of the audio itself, we think it may have better results, which is also the direction we want to continue exploring.

Table 11. Comparative Performance of Different Methods on the VGGSound Dataset.

Method	Model Param	Pre-Trained Data	VGGSound
VGGSound [27]	81M	-	0.488
CAV-MAE [45]	87M	ImageNet + AudioSet	0.595
MBT [12]	87M	ImageNet 21K	0.523
Aud-SlowFast [46]	-	-	0.501
MAViL [25]	87M	ImageNet + AudioSet	0.608
ACM-SSKD (iter1)	92M	ImageNet + AudioSet	0.579
ACM-SSKD (iter2)	92M	ImageNet + AudioSet	0.588
ACM-SSKD (iter3)	92M	ImageNet + AudioSet	0.605
ACM-SSKD (Ensemble)	92M	ImageNet + AudioSet	0.613

Table 12. Comparative Performance of Different Methods on the ESC-50 Dataset.

Method	Model Param	Pre-Trained Data	ESC-50
PANN [1]	81M	-	0.947
AST [5]	86M	ImageNet	0.956
ERANN [14]	55M	AudioSet	0.961
Audio-MAE [19]	86M	AudioSet	0.974
Ssast [24]	89M	AudioSet + LibriSpeech	0.888
MaskedSpec [18]	86M	AudioSet	0.896
Mae-ast [19]	86M	AudioSet + LibriSpeech	0.900
SPFA [44]	87M	-	0.968
BEATS (iter3) [21]	90M	AudioSet	0.956
BEATS (iter3+) [21]	90M	AudioSet	0.981
ACM-SSKD (iter1)	92M	ImageNet + AudioSet	0.962
ACM-SSKD (iter2)	92M	ImageNet + AudioSet	0.975
ACM-SSKD (iter3)	92M	ImageNet + AudioSet	0.984
ACM-SSKD (Ensemble)	92M	ImageNet + AudioSet	0.987

Table 11 displays the comparative results of our method on the VGGSound dataset against other methods. The experimental setup aligns with that of the audioset experiment. As our approach is based on modeling the audio single modality, the results demonstrate a noticeable improvement compared to VGGSound [27], CAV-MAE [31], MBT [12], and Aud-SlowFast [46]. However, even in comparison with the audio-visual multimodal approach of MAViL [25], our single-model approach exhibits a recognition difference of 0.3% in iter3, highlighting the effectiveness of joint modeling of audio-visual multimodalities, an optimization direction for future exploration. Nevertheless, through ACM-SSKD (ensemble) joint recognition, our paper achieves the best results, further confirming the effectiveness of joint recognition.

Table 12 presents a comparative analysis of our method on the ESC-50 dataset against other methods. Similarly, AST [5] was selected as the knowledge distillation teacher model. The results reveal that under the same number of iterations, our approach outperforms the BEATS method in recognition metrics. Following ensemble result fusion, there is a further enhancement in recognition results.

5. Conclusions

This paper, grounded in the context of audio single modality, explores a novel framework that integrates self-supervision with knowledge distillation to enhance the performance of audio classification tasks. In the realm of self-supervision, we construct temporal and frequency domain random masks, coupled with spectrogram block random masks for information reconstruction. This approach facilitates effective learning of intricate details and crucial discriminative information in audio spectra through contrastive learning and feature distribution fitting. To enhance training efficiency, knowledge distillation is employed to emulate the feature expression capabilities of a teacher model.

Multiple ablation experiments were conducted on publicly available datasets, including intra-patch spectrograms and feature reconstruction, inter-patch spectrograms and feature reconstruction, multidimensional self-supervision, pre-training weight loading, and audio knowledge distillation based on teacher models. The results indicate that the proposed ACM-SSKD algorithmic framework, as presented in this paper, through a multidimensional self-supervised strategy based on audio spectrograms combined with teacher distillation can effectively learn and distinguish complex audio features. Advanced results have been achieved in pure audio recognition on three publicly available datasets, AudioSet-2M, ESC-50, and VGG Sound.

Although this paper attains commendable results in audio classification, it acknowledges the associated challenges of increased training costs and operational complexity in the two-stage training process. Simultaneously, our research reveals that exploring the realm of audio-visual multimodal joint modeling holds promising implications for advancing the effectiveness of audio classification. These two aspects serve as directions for ongoing optimization and improvement in future research endeavors.

Author Contributions: X.G. data curation, conceived and designed the experiments; H.D. and Y.Y. formal analysis; L.T., J.W. and A.V.V. writing review and supervison. All authors have read and agreed to the published version of the manuscript.

Funding: This work is partially supported by the National Natural Science Foundation of China under Grant 62173345, partially supported by the Natural Science Foundation of Shandong Province under Grants ZR2022LZH015 and 2023QF025, partially supported by the Integrated Innovation of Science, Education and Industry of Qilu University of Technology (Shandong Academy of Sciences) under Grant 2023PX057, and partially supported by the Talent Project of Qilu University of Technology (Shandong Academy of Sciences) under Grant 2023RCKY141.

Data Availability Statement: The Sample data used to support the findings of this study are available at https://research.google.com/audioset/download.html, https://github.com/karoldvl/ESC-50/archive/master.zip, http://www.robots.ox.ac.uk/~vgg/data/vggsound/ (accessed on 17 December 2023).

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

Appendix A

In our experiments, one-dimensional audio was initially transformed into two-dimensional spectrograms. For a T-second audio signal, computations were performed with a 25ms calculation window and a 10ms step size. The quantized feature dimension was set to 128, resulting in a two-dimensional spectrogram feature. Following these operations, a 10-s audio yielded a resolution of 1024 × 128 for the two-dimensional spectrogram, while a 5-s audio produced a resolution of 512 × 128. Similar to the approach in VIT [15], the two-dimensional audio spectrogram is decomposed into non-overlapping 16 × 16 image patches. Considering the practical significance of the two-dimensional spectrogram in time and frequency, this can be understood as sampling with a step size of 16 in both dimensions. Consequently, for a 10-s video, a sequence length of 512 and a feature length of 256 were obtained as input; for a 5-s video, a sequence length of 256 and a feature length of 256 were obtained.

When implement the pre-training algorithm in this paper, the AudioSet 2M dataset is used. Similar to the AST [5] and BEATS[q0] models, the ACM-SSKD model has 12 encoder layers based on the Transformer, with a hidden state dimension of 768 and 12 attention heads. Adding two layers of perceptrons (768, 1536) and (1536, 768) to the last layer, and the total model parameter approximately 92 M.

We use 32 V100 GPUs for training, and the dataset we used is AudioSet-2M. At each iteration, the batch size is 256, and the learning rate is 0.00015. After each iteration, the learning rate decayed by 20%. The total training time for each iteration is approximately 28 h, which includes multidimensional self-supervision and distillation based on teacher models. For each audio sample, we randomly sample the start time and loop to extract 10 s of audio. And the Algorithm A1 describes the training process of this article using pseudocode.

Algorithm A1 ACM-SSKD Pseudocode of this paper in a PyTorch-like style

```
# spectrograms: Convert audio signals into two-dimensional spectrogram image.
# intra_patch, patch, inter_patch: Blocking and masking operations.
# AFT: Audio feature Transformer, as shown in Figure 2.
# MAE: Audio mask auto encoder, as shown in Figure 2.
for iter in iters: # iters represents the number of iterations
for x in batch: #x is one-dimensional audio raw data
  x_spec = spectrograms(x) # x_spec is two-dimensional spectrogram image
  x_raw = patch(x_spec) # patch raw, as shown in Figure 2
  x_imtra = intra_patch(x_raw) # intra-patch mask, as shown in Figure 2
  x_inter = intra_patch(x_raw) # intra-patch mask, as shown in Figure 2
# Multi-SSL process, res * and fea *, respectively, represent reconstructed images
# and features extracted through networks
  res_raw, fea_raw = AFT(x_raw)
  res_intra, fea_intra = AFT(x_intra)
  res_inter, fea_inter = AFT(MAE(x_inter))
  #restore spectrogram image, WMse as shown in Formula (3).
  l_spec_res = WMse(res_intra, res_raw) + WMse(res_inter, res_raw)
  # restore the features generated from the patch raw, MSE as shown in Formula (6).
  l_fea_mse = Mse(fea_intra, fea_raw) + Mse(fea_inter, fea_raw)
  # restore the features generated from the patch raw, KL as shown in Formula (7).
  l_fea_kl = KL(fea_intra, fea_raw) + KL(fea_inter, fea_raw)
  # comparative learning of features, Cont as shown in Formula (9).
  l_fea_cont = Cont(AVG(fea_intra), AVG(fea_raw)) + Cont(AVG(fea_inter), AVG(fea_raw))
  l_multi_ssl = $\alpha \times$ (l_spec_res + l_fea_mse)+ $\beta \times$ l_fea_kl + l_fea_cont
  #end for Multi-SSL

  # Knowledge distillation process.
  # CE as shown in Formula (11), Reg as shown in Formula (12).
  # fea_cls is generated by AFT network after Multi-SSL operate.
  # fea_teacher is generated by teacher model.
  l_cls = CE(fea_cls, fea_teacher)
  l_res = Mse(fea_cls, fea_teacher)
  l_kd = l_cls + $\gamma \times$ l_res
  # end for Knowledge distillation
```

References

1. Kong, Q.; Cao, Y.; Iqbal, T.; Wang, Y.; Wang, W.; Plumbley, M.D. Panns: Large-scale pretrained audio neural networks for audio pattern recognition. *IEEE ACM Trans. Audio Speech Lang. Process.* **2020**, *28*, 2880–2894. [CrossRef]
2. Hsu, W.N.; Bolte, B.; Tsai, Y.H.H.; Lakhotia, K.; Salakhutdinov, R.; Mohamed, A. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3451–3460. [CrossRef]
3. Verma, P.; Berger, J. Audio Transformers: Transformer Architectures for Large Scale Audio Understanding. In Proceedings of the IEEE Workshop on Applications of Signal Processing to Audio and Acoustics, New Paltz, NY, USA, 17–20 October 2021; pp. 1–5.
4. Arnault, A.; Hanssens, B.; Riche, N. Urban Sound Classification: Striving towards a fair comparison. *arXiv* **2020**, arXiv:2010.11805.

5. Gong, Y.; Chung, Y.A.; Glass, J. AST: Audio Spectrogram Transformer. In Proceedings of the IEEE Conference on Interspeech, Brno, Czechia, 30 August–3 September 2021; pp. 571–575.
6. Liu, A.T.; Li, S.W.; Tera, H.L. Self-supervised learning of transformer encoder representation for speech. *IEEE ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 2351–2366. [CrossRef]
7. Chi, P.H.; Chung, P.H.; Wu, T.H.; Hsieh, C.C.; Chen, Y.H.; Li, S.W.; Lee, H.Y. Audio albert: A lite bert for self-supervised learning of audio representation. In Proceedings of the IEEE Spoken Language Technology Workshop (SLT), Shenzhen, China, 19–22 January 2021; pp. 344–350.
8. Tan, M.; Le, Q. Efficientnet: Rethinking model scaling for convolutional neural networks. In Proceedings of the 36th International Conference on Machine Learning, Long Beach, CA, USA, 9–15 June 2019; pp. 6105–6114.
9. Giraldo, J.S.P.; Jain, V.; Verhelst, M. Efficient Execution of Temporal Convolutional Networks for Embedded Keyword Spotting. *IEEE Trans. Very Large Scale Integr. (VLSI) Syst.* **2021**, *29*, 2220–2228. [CrossRef]
10. Yuan, G.; Yu, A.C.; James, G. PSLA: Improving Audio Tagging with Pretraining, Sampling, Labeling, and Aggregation. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2021**, *29*, 3292–3306.
11. Schmid, F.; Koutini, K.; Widmer, G. Efficient Large-Scale Audio Tagging Via Transformer-to-CNN Knowledge Distillation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.
12. Arsha, N.; Shan, Y.; Anurag, A.; Jansen, A.; Schmid, C.; Sun, C. Attention bottlenecks for multimodal fusion. *J. Adv. Neural Inf. Process. Syst.* **2021**, *34*, 14200–14213.
13. 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 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 646–650.
14. Sergey, V.; Vladimir, B.; Viacheslav, V. Eranns: Efficient residual audio neural networks for audio pattern recognition. *J. Pattern Recognit. Lett.* **2022**, *161*, 38–44.
15. Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; Jégou, H. Training data-efficient image transformers & distillation through attention. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 8–24 July 2021; pp. 10347–10357.
16. Ze, L.; Yutong, L.; Yue, C.; Han, H.; Wei, Y.; Zheng, Z.; Stephen, L.; Baining, G. Swin transformer: Hierarchical vision transformer using shifted windows. In Proceedings of the IEEE/CVF International Conference on Computer Vision, Montreal, QC, Canada, 10–17 October 2021; pp. 10012–10022.
17. Xie, X.; Zhang, H.; Wang, J.; Chang, Q.; Wang, J.; Pal, N.R. Learning optimized structure of neural networks by hidden node pruning with L1 regularization. *IEEE Trans. Cybern.* **2020**, *50*, 1333–1346. [CrossRef]
18. Dading, C.; Helin, W.; Peilin, Z.; Zeng, Q.C. Masked spectrogram prediction for self-supervised audio pre-training. *arXiv* **2022**, arXiv:2204.12768.
19. Huang, P.Y.; Xu, H.; Li, J.; Baevski, A.; Auli, M.; Galuba, W.; Metze, F.; Feichtenhofer, C. Masked autoencoders that listen. *arXiv* **2022**, arXiv:2207.06405.
20. Yu, Z.; Daniel, S.P.; Wei, H.; Qin, J.; Gulati, A.; Shor, J.; Jansen, A.; Xu, Y.Z.; Huang, Y.; Wang, S. Bigssl: Exploring the frontier of large-scale semi-supervised learning for automatic speech recognition. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1519–1532.
21. Chen, S.; Wu, Y.; Wang, C.; Liu, S.; Tompkins, D.; Chen, Z.; Wei, F. BEATS: Audio Pre-Training with Acoustic Tokenizers. In Proceedings of the 40th International Conference on Machine LearningJuly, Honolulu Hawaii, HI, USA, 23–29 July 2023; pp. 5178–5193.
22. Baevski, A.; Hsu, W.N.; Xu, Q.; Babu, A.; Gu, J.; Auli, M. Data2vec: A General Framework for Self-supervised Learning in Speech, Vision and Language. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; pp. 1298–1312.
23. Chen, S.; Wang, C.; Chen, Z.; Wu, Y.; Liu, S.; Chen, Z.; Li, J.; Kanda, N.; Yoshioka, T.; Xiao, X.; et al. Wavlm: Large-scale self-supervised pre-training for full stack speech processing. *IEEE J. Sel. Top. Signal Process.* **2022**, *16*, 1505–1518. [CrossRef]
24. Gong, Y.; Lai, C.I.; Chung, Y.A.; Glass, J. Ssast: Self-supervised audio spectrogram transformer. In Proceedings of the AAAI Conference on Artificial Intelligence, Vancouver, BC, Canada, 22 February–1 March 2022; pp. 10699–10709.
25. Huang, P.Y.; Sharma, V.; Xu, H.; Ryali, C.; Fan, H.; Li, Y.; Li, S.W.; Ghosh, G.; Malik, J.; Feichtenhofer, C. MAViL: Masked Audio-Video Learners. *arXiv* **2022**, arXiv:2212.08071.
26. Wei, Y.; Hu, H.; Xie, Z.; Zhang, Z.; Cao, Y.; Bao, J.; Chen, D.; Guo, B. Contrastive learning rivals masked image modeling in fine-tuning via feature distillation. *arXiv* **2022**, arXiv:2205.14141.
27. Chen, H.; Xie, W.; Vedaldi, A.; Zisserman, A. VGGSound: A large-scale audio-visual dataset. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Barcelona, Spain, 4–8 May 2020; pp. 721–725.
28. Wei, C.; Fan, H.; Xie, S.; Wu, C.Y.; Yuille, A.; Feichtenhofer, C. Masked feature prediction for self-supervised visual pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 14668–14678.
29. Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Taylor, B.K.; Dubnov, S. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, Rhodes Island, Greece, 4–10 June 2023; pp. 1–5.

30. Aaqib, S.; David, G.; Neil, Z. Contrastive learning of general-purpose audio representations. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 3875–3879.
31. Eduardo, F.; Diego, O.; Kevin, M.; Noel, E.O.C.; Serra, X. Unsupervised contrastive learning of sound event representations. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 371–375.
32. Haider, A.T.; Yalda, M. Clar: Contrastive learning of auditory representations. In Proceedings of the International Conference on Artificial Intelligence and Statistics, Toronto, ON, Canada, 6–11 June 2021; pp. 2530–2538.
33. Luyu, W.; Aaron, O. Multi-format contrastive learning of audio representations. *arXiv* **2021**, arXiv:2103.06508.
34. Daisuke, N.; Daiki, T.; Yasunori, O.; Harada, N.; Kashino, K. Byol for audio: Self-supervised learning for general-purpose audio representation. In Proceedings of the 2021 International Joint Conference on Neural Networks (IJCNN), Shenzhen, China, 18–22 July 2021; pp. 1–8.
35. Alan, B.; Puyuan, P.; David, H. Mae-ast: Masked autoencoding audio spectrogram transformer. In Proceedings of the 23rd Interspeech Conference, Incheon, Republic of Korea, 18–22 September 2022; pp. 2438–2442.
36. Andrew, N.C.; Quentin, B.; Mathieu, B.; Teboul, O.; Zeghidour, N. Selfsupervised learning of audio representations from permutations with differentiable ranking. *J. IEEE Signal Process. Lett.* **2021**, *28*, 708–712.
37. Gong, X.; Yu, L.; Wang, J.; Zhang, K.; Bai, X.; Pal, N.R. Unsupervised Feature Selection via Adaptive Autoencoder with Redundancy Control. *Neural Netw.* **2022**, *150*, 87–101. [CrossRef]
38. Aditya, R.; Mikhail, P.; Gabriel, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; Sutskever, I. Zero-shot text-to-image generation. In Proceedings of the 38th International Conference on Machine Learning, Virtual, 8–24 July 2021; pp. 8821–8831.
39. Bao, H.; Dong, L.; Piao, S.; Wei, F. Beit: Bert pre-training of image transformers. *arXiv* **2021**, arXiv:2106.08254.
40. Feichtenhofer, C.; Li, Y.; He, K. Masked Autoencoders as Spatiotemporal Learners. *arXiv* **2022**, arXiv:2205.09113.
41. Liu, A.T.; Yang, S.; Chi, P.H.; Hsu, P.C.; Lee, H. Mockingjay: Unsupervised speech representation learning with deep bidirectional transformer encoders. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 6419–6423.
42. Gemmeke, J.F.; Ellis, D.P.W.; Freedman, D.; Jansen, A.; Lawrence, W.; Moore, R.C.; Plakal, M.; Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 776–780.
43. 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; pp. 1015–1018.
44. Gong, X.; Li, Z. An Improved Audio Classification Method Based on Parameter-Free Attention Combined with Self-Supervision. *J. Comput.-Aided Des. Comput. Graph.* **2023**, *35*, 434–440.
45. Yuan, G.; Andrew, R.; Alexander, H.L.; Harwath, D.; Karlinsky, L.; Kuehne, H.; Glass, J. Contrastive Audio-Visual Masked Autoencoder. In Proceedings of the 17th International Conference on Learning Representations, Kigali, Rwanda, 1–5 May 2023; pp. 1–29.
46. Evangelos, K.; Arsha, N.; Andrew, Z.; Dima, D. Slow-fast auditory streams for audio recognition. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 855–859.

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.