MASKED MODELING DUO VISION TRANSFORMER WITH MULTI-LAYER FEATURE FUSION ON RESPIRATORY SOUND CLASSIFICATION

Boxin Liu[1], *Shiqi Zhang*[1], *Daiki Takeuchi*[2], *Daisuke Niizumi*[2], *Noboru Harada*[2], *Shoji Makino*[1]

[1]Waseda University, Japan [2] NTT Corporation, Japan

ABSTRACT

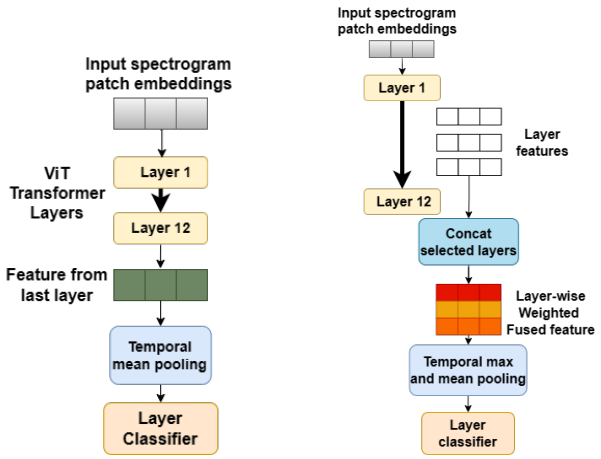
Respiratory sounds are significant relevant indicators for respiratory health and conditions. Classifying the respiratory sounds of patients can assist doctors' diagnosis of lung diseases. For this purpose, many deep learning-based automatic analysis methods have been developed. However, it is still challenging due to the limited medical sound datasets. In this study, we apply a pre-trained Vision Transformer (ViT) based model from the Masked Modeling Duo (M2D) framework for this task. While the M2D ViT pre-trained model provides effective features, we think combining features from different layers can improve the performance in this task. We propose a multi-layer feature fusion method using learnable layer-wise weights and validate its effectiveness in experiments and an analysis of pre-trained model layers. Our approach achieves the best ICBHI score of 60.68, 2.39 higher than the previous state-of-the-art method.

Index Terms— Respiratory Sound Classification, ICBHI, Pre-trained Model, Feature Fusion, Masked Modeling Duo

1. INTRODUCTION

Respiratory diseases have recently become the third cause of death worldwide [1]. And due to the impact of the COVID-19 global pandemic, the need for diagnosing lung disease with efficient methods with accuracy and lower work burden for physicians and medical experts has been increasing. Respiratory sound classification is a task to identify whether a breathing cycle of a recorded sound sample contains adventitious sounds related to potential disease in the respiratory system. Conventional respiratory sound classification requires medical experts to utilize stethoscopes to conduct auscultations for patients in person, which is highly demanding for hospitals and other medical institutions [2].

International Conference on Biomedical Health Informatics (ICBHI) Respiratory Sound Database [3] is a public database for developing the algorithms on respiratory classification tasks recorded by microphones and electronic stethoscopes. The audio samples in this dataset consist of respiratory cycles in variant lengths with four kinds of annotations: normal, crackles, wheeze, and the combination of both anomalies. Crackles are discontinuous adventitious sounds in breathing cycles and can be an early sign of cardiorespiratory conditions. At the same time, wheezes are continuous and musical sounds of anomaly, indicating the patient's obstructive airway conditions. The classification for these types of breath sounds can be the basis for diagnosing or monitoring diseases such as asthma, Chronic Obstructive Pulmonary Disease (COPD) [4], and pneumonia. With the release of this dataset, more and more research attention has been drawn to the respiratory sound classification task and further the automatic assistance for doctors' diagnoses.



(a) Conventional (b) Proposed

Figure 1: Overview of the conventional method and the multi-layer feature fusion workflow.

The sounds in the ICBHI dataset were recorded from various positions and by many types of equipment, making distinguishing different respiratory cycles difficult. Besides, the dataset scale is limited. Until now, several previous studies have proposed models to tackle this task, and many novel structures or algorithms and data augmentations have been introduced [5–11]. With the addition of the limited dataset, pre-training visual models with large-scale datasets have been widely used in ICBHI task [7–11].

The models from the self-supervised learning framework pre-training on large-scale audio datasets have recently achieved competitive performance in the image field and several audio tasks [12–15]. In this study, we used pre-trained ViT [16] models from the Masked Modeling Duo (M2D) framework [15] (M2DViT). We adopted the M2DViT without changing the backbone configurations, such as patch size and grid size.

While the features from the M2DViT can perform well in various tasks, we expect to achieve even better performance by combining features from different layers to form effective representations of audio samples. In this study, we explore the possibility of the feature fusion available from the M2DViT layers for solving the ICBHI task and propose methods for fusing effective features. We experiment with our methods on the ICBHI task and validate the effectiveness. In addition, we analyze the contribution of layer features and show that the later layers contribute more.

In summary, the main contributions of this paper are as follows:

- Proposing to compose effective representations for the respiratory sound classification task using M2D layer features.

- Conducting experiments with our multi-layer feature fusion methods and comparing ours with the previous methods.
- Analyzing the performance of different layers of M2DViT and their combinations in the respiratory sound classification task.

2. RELATED WORKS

2.1. Respiratory Sound Classification

Ever since the ICBHI2017 challenge and the release of the open access dataset, researchers have trained and evaluated many deep learning-based respiratory classifying methods to have better solutions for this task. Early models like LungRN+NL [5] combine ResNet-based architecture and mix-up augmentation method, and then the attention mechanism was introduced with LungAttn [6]. The works after 2020 are widely presented with ImageNet [17] or AudioSet [18] pre-training. And for RespireNet [8], the authors also use a device-specified fine-tuning strategy to improve the performance. The previous works are mainly based on ResNet structure except for a recent work [11], which uses a simple CNN backbone from PANNs [19] and contrastive learning with metadata strategy. The self-supervised methods, such as contrastive learning, show their validity in [11]. A concurrent work based on Audio Spectrogram Transformer (AST) [20] and contrastive learning with PatchMix augmentation [21] shows that the pre-trained attention-based model has the potential for better performance than other conventional models.

2.2. Masked Modeling Duo

The adopted self-supervised learning framework of M2D [15] is an effective method for general-purpose pre-training using a masked prediction task. This method was originally inspired by the Masked Autoencoder (MAE) [22] approach utilized in Masked Image Modeling (MIM), along with the Bootstrap Your Own Latent (BYOL) [23] framework, which enables the direct acquisition of latent representations through a target network.

In the two divided networks of M2D, the framework learns to predict the output of the target network with the output of the online network. At the same time, visible patches serve as input for the online network, and masked patches for the target network. While the online network weights are optimized to minimize the loss, the weights of the encoder in target network ξ are updated based on the exponential moving average (EMA) of the online network θ with a decay rate τ.

M2D learns effective representations in the online encoder. After training, only the trained parameter of the online encoder f_θ is transferred as a pre-trained ViT model, which we call M2DViT, for downstream tasks. The M2DViT pre-trained weights are available online[1] and used in our experiments, which are pre-trained on AudioSet [18]. Unlike previous works, we combine multi-layer feature outputs.

2.3. Feature Fusion

The method of Feature Fusion was broadly proposed to deal with multi-modal tasks [24, 25]. There are numerous approaches to extracting features from different levels of deep learning models. The skip connection structure and multi-scale attention mechanism have been widely used. For example, in the work, MS-CAM [26], an iterative attentional feature fusion method performs excellently in vision models. And for another instance, in the work of MFVT [27], the authors proved that the fused features in the ViT-based model are a potent strategy in the fine-grained visual categorization task. Besides, in audio-related tasks, the multi-layer feature fusion serves as a powerful method, as reported in [28]. The mechanism of multi-layer feature fusion is similar to the skip connection essential for convolutional networks such as ResNet [29] and DenseNet [30], and various methods for connecting layers are proposed. The skip connections encourage the networks to obtain semantic features from the early layers of the model [31]. The fusion is usually performed by operations of addition or concatenation with a fixed weight of the features [29, 30].

[1] https://github.com/nttcslab/m2d

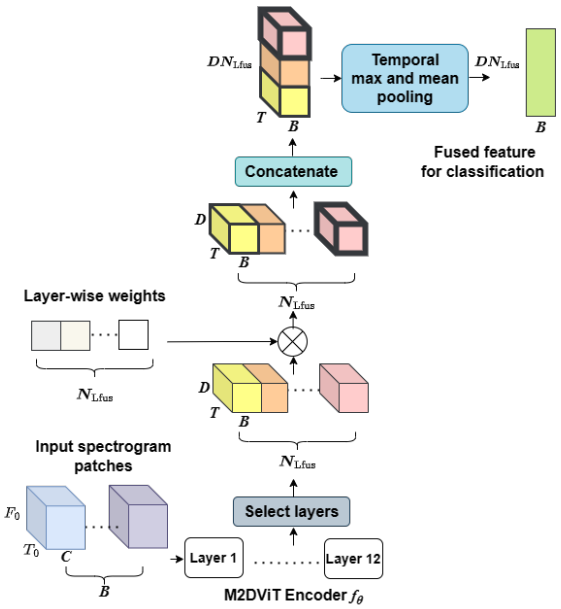


Figure 2: The multi-layer feature fusion calculation flow. F_0, T_0, and C are frequency bins, time frames, and channels in a spectrogram, respectively.

3. METHODOLOGY

The encoder in M2DViT is based on the ViT backbone, consisting of 12 transformer blocks as layers. The ViT first patchifies the input mel-spectrogram and then processes it with a projection of a linear layer, transferring the spectrogram into patch embeddings. Then the fixed sinusoidal positional embedding is added to the input. The multi-head attention is applied , followed by the MLP containing 2 linear layers with a Gaussian error linear unit (GELU) activation. We denote the transformer blocks as transformer layers for simplicity. The outputs from all transformer layers have the same shape, and all the layer outputs are available for later use, such as classification.

The conventional M2DViT, shown in Fig. 1(a), takes a spectrogram input, processes the input in the transformer layers, then outputs the last layer feature $z \in R^{B \times T \times D}$, where B is the input batch size, T is the length of the sequence composed by encoded spectrotemporal patches, and D is the embedded patch feature dimension. Then, only the output z is used afterward.

One drawback of the conventional method is that the information from different layers is not used as the representation for audio [32]. The performance of the features from pre-trained ViT layers can be imbalanced due to the structure and training metrics [28]. To address this problem, we introduce multi-layer feature fusion methods to combine the layer features. We also use learnable layer-wise weights to balance layers with better performance automatically, which is optimized as the training epoch proceeds. Our approach enables arbitrary combinations of the layer features as effective representations for later use.

The pipeline for multi-layer feature fusion is shown in Fig. 1(b). And Fig. 2 shows the details of the feature calculation flow. The output features of all the layers can be defined as $\left\{z_i \in R^{B\times T\times D} | i \in L\right\}$, where z_i is the i-th layer output and L is the number of layers.

Then, we calculate the multi-layer feature fusion $\tilde{z}$ as follows:

$$\tilde{z} = \text{concat}(\{w_i z_i | i \in L_{\text{fus}}\}) \tag{1}$$

where the concat is a function that concatenates features on the dimension of D, L_{fus} is the set of layer indexes of desired fusion, and w_i is a learnable layer-wise weight of the i-th layer in the fused feature. Note that the $\tilde{z}$ forms the shape of $R^{B\times T\times D N_{\text{Lfus}}}$, where N_{Lfus} is the number of the fused layers. As a result, the multi-layer feature fusion enables us to utilize useful information in the features from all the desired layers.

As a final operation, we apply temporal poolings to summarize time-framed features in a feature vector:

$$z' = \text{mean}(\tilde{z}) + \max(\tilde{z}) \tag{2}$$

where $z' \in R^{B\times D N_{\text{Lfus}}}$ is the final fused feature vector used as the input for later use (e.g., classification) and mean/max are temporal operations each. We follow [28] for the effective temporal pooling operation.

4. EXPERIMENTS

We conducted experiments to validate our proposals. The following sections explain the dataset (Section 4.1), evaluation metrics (Section 4.2), and experimental setup (Section 4.3). Then we show experimental results with vanilla M2DViT (Section 4.4) and results with our proposals as well as previous studies (Section 4.5).

4.1. Dataset

ICBHI Respiratory Sound Database [3] consists of 920 annotated respiratory audio samples recorded from 126 patient subjects of the labs and hospitals in Portugal and Greece. The samples are officially split into a train set (539 samples, 60%) and a test set (381 samples, 40%). The database contains two sets of annotations. One is for whether a cycle contains crackles, wheezes, or a combination of both, and some with no adventitious respiratory sounds. The other is the annotation of the locations of the adventitious respiratory sounds. In the 6898 respiratory cycles, whose lengths vary from 0.2s to 16.2s, 1864 contain crackles, 886 contain wheezes, and 506 contain both crackles and wheezes. The others are normal ones. The chest locations from which the recordings were acquired are also provided. Noise levels in some respiration cycles are high, which simulates real-life conditions. The recordings were collected using heterogeneous equipment, and their duration ranged from 10s to 90s. The average time duration of the cycles is 2.7s, and the total length is 5.5h.

Method	Masking ratio r	S_p	S_e	S_c
M2DViT	0.6	71.59±2.64	43.25±1.60	57.42±0.79
	0.7	75.78±5.84	39.51±4.64	57.64±0.86

Table 1: The ICBHI performance comparison between different pre-training masking ratios of M2DViT.

4.2. Evaluation Metrics

The evaluation metrics in our experiments are adopted from the original ICBHI2017 challenge, which is common in the previous papers. There are three scores, sensitivity S_e, specificity S_p, and the average of these two metrics ICBHI score S_c. They are calculated as the following formulas:

$$S_e = \frac{P_c + P_w + P_b}{T_c + T_w + T_b} \tag{3}$$

$$S_p = \frac{P_n}{T_n} \tag{4}$$

$$S_c = \frac{S_e + S_p}{2} \tag{5}$$

where P_c, P_w, P_b, and P_n are numbers of right prediction for the cycles containing crackles, wheezes, both of the two adventitious sounds and none of them. While T_c, T_w, T_b, and T_n are the total numbers of four categories respectively.

4.3. Experimental Setup

We used an Adam optimizer with a learning rate of 1e-4 and weight decay of 1e-4, cosine scheduled in the M2DViT model. The batch size is set as 64. The classifier used in our study is a 4-class linear classifier. The input spectrograms are patchified with a patch size of (16, 16), and the grid size is (5, 38). The number of encoder embedding dimensions is 768. The ViT-base pre-trained model consists of 12 transformer blocks with the same number of attention heads [16]. We fine-tuned all the pre-trained ViT model weights in 150 epochs and used weighted cross-entropy as our training and evaluating loss.

We used the same ViT and training settings described above for all the setups. For the adaptive weight for all layers, we initialized them as all layers weighted the same value of 1.0 and updated them after training every epoch. All our experiments run five times with random seeds, and we provide statistics of results.

4.3.1. Preprossessing and Data Augmentation

We followed the basic experiments settings with [11], in which the authors also used mel-spectrogram as input. Due to the dataset's recording conditions, the audio samples' sampling rates vary in an extensive range from 4 kHz to 44.1 kHz. We first resampled them into a fixed sampling rate of 16kHz. And for different durations of the samples, we ensure that all samples have the same desirable length of 8s. For longer samples, we limited them to 8s from the beginning of each clip. While for the shorter samples, we circularly pad them until we get the standard length. Because most of the respiratory cycles are shorter than 8s.

In this length of time, the model can compose representations for most of the respiratory cycles. The spectrogram transform settings in our experiments are the default in M2D. The samples are converted into a time-frequency representation of a log-mel spectrogram with 80 mel filterbanks, a window length of 400, and a hop length of 160. The minimum and maximum frequencies are 50 Hz

Method	Architecture	Pre-training dataset	Fused Layers	Layer-wise weights	Results S_p	S_e	S_c
LungRN+NL [5]	ResNet	-	-	-	63.20	41.32	52.26
LungAttn [6]	ResNet	-	-	-	71.44	36.36	53.90
Wang et al. [7]	ResNeSt [33]	ImageNet	-	-	70.40	40.20	55.30
RespireNet [8]	ResNet34	ImageNet	-	-	72.30	40.10	56.20
ARSC-Net [9]	ResNet	-	-	-	67.13	**46.38**	56.76
Nguyen et al. [10]	ResNet50	ImageNet	-	-	79.34	37.24	58.29
Moummad et al. [11]	CNN6 [19]	AudioSet	-	-	75.95	39.15	57.55
M2DViT	M2D ViT [15, 16] masking ratio=0.7	AudioSet	-	-	75.78±5.84	39.51±4.64	57.64±0.86
M2DViT-Fusion			(i) 5th & 12th	Fixed (1.0 for all)	75.43±5.22	41.18±5.80	58.30±0.97
			(ii) 5th & 11th	Fixed (1.0 for all)	**82.05±4.16**	38.06±3.24	60.05±1.00
			(iii) 4th & 7th & 10th	Fixed (1.0 for all)	79.69±2.68	39.96±1.56	59.83±0.72
			(iv) All	Fixed (1.0 for all)	79.71±3.58	40.34±2.55	59.97±0.69
			(v) All	Learnable	79.48±4.99	41.87±4.27	**60.68±0.49**

Table 2: The overall comparison of ICBHI performance of our methods and the previous ones. Except for the last result with learnable layer-wise weight, all fused layer features are of the fixed weight of 1.0. All the results are presented with the mean values and standard deviations.

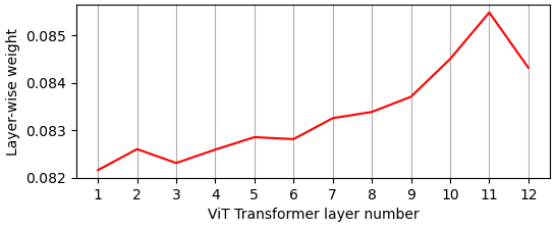


Figure 3: The learned layer-wise feature weights of the M2DViT-Fusion model with the best ICBHI score S_c. The weights are the average of best epochs from 5 runs and normalized to the sum of 1.

and 8000 Hz. Then the transformed spectrogram is normalized and standardized into a mean value of 0.3690 and a standard deviation of 0.2550. We also used the augmentation method SpecAugment [34] as in [11]. The mask sizes for time and frequency are 20 and 40, while stripes are 2 for both time and frequency in SpecAugment. We shuffled the train samples, masking the blocks of the frequency and time steps with time-wrapping augmentation to encourage the network to learn robust features from the spectrogram.

4.4. Experiments with Vanilla M2DViT

We first compared M2DViT weights pre-trained with masking ratios of 0.6 and 0.7 by fine-tuning them without our proposals. Table 1 shows that pre-trained M2DViT with a 0.7 masking ratio (M2D r=0.7) performs slightly better than M2D r=0.6, and the S_c of these two options are almost identical. While the original M2D r=0.6 showed better results on S_e, M2D r=0.7 weight was better on S_p. The original M2D r=0.6 showed better results on the speech tasks, and M2D r=0.7 was better on music tasks [15]. ICBHI with respiratory sounds is supposed to be more speech-like breathing noise [3]. Therefore, we used M2D r=0.7 in the following experiments.

4.5. Experiments with Proposals

We applied our multi-layer feature fusion methods with various layer combinations, denoted as M2DViT-Fusion, and compared them with the previous methods. Table 2 shows the results of the best-performing layer combinations in the brute-force parameter search, and Fig. 3 shows the learned layer-wise weights in the M2DViT-Fusion of the epoch with the best ICBHI score S_c. It is worth noting that the representation dimensionality varies from D in M2DViT to $D \times 2$ in (i) and (ii), $D \times 3$ in (iii), and $D \times 12$ in (iv) and (v). For the sake of experiment time constraints, we learned layer-wise weights only when using all layers.

We find that the 11th layer would provide the most significant features for the ICBHI task; Fig. 3 shows that the 11th is the best for (v) in Table. 2 with all layer fusion with learnable layer-wise weights, and the (ii) 5th & 11th layer fusion shows the best S_p result. Fig. 3 also shows a trend that the later layers perform better, though the performance drops at the last layer.

We also found that learning the layer-wise weights is better than the fixed weights, showing the effectiveness of the layer feature weighting; while the results of (v) with learnable weights and (iv) with fixed weights are highly overlapping.

Compared with the previous studies, (v) fusing all layers with learnable weights shows the best average ICBHI score of 60.68±0.49, about 2 point improvement from Nguyen et al. [10], with a score of 58.29. For the S_p, (ii) 5th & 11th shows the best result of 82.05±4.16, more than 2 point improvement from Nguyen et al., with a score of 79.34. However, for the S_e, ARSC-Net shows the best result of 46.38. Overall, we think the results validated the effectiveness of our proposals.

5. CONCLUSION

We introduced a novel feature fusion method to the classification task on the ICBHI dataset. And in the experiments, our M2DViT-Fusion methods showed a better performance than the vanilla M2DViT. The results proved that multi-layer feature fusion is an effective way to extract effective audio representations, including the proposed learnable layer-wise weight. In the layer weight analysis, we also found the later layers contribute more. The fine-tuned model got the best ICBHI score of 60.68 on the ICBHI dataset, which is improved by 2.39 compared to the previous SOTA method. While we exhibited improvements, the result would still need further improvement for practical diagnosis assistance. Possible directions may include effective augmentation techniques and new large-scale respiratory sound datasets to help models achieve desirable performance in the future.

6. ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number 23H03423.

7. REFERENCES

[1] S. M. Levine and D. D. Marciniuk, "Global impact of respiratory disease: What can we do, together, to make a difference?" *Chest*, vol. 161, no. 5, pp. 1153–1154, 2022.

[2] M. A. Fernandez-Granero, D. Sanchez-Morillo, and A. Leon-Jimenez, "Computerised analysis of telemonitored respiratory sounds for predicting acute exacerbations of COPD," *Sensors*, vol. 15, no. 10, pp. 26 978–26 996, 2015.

[3] B. M. Rocha, D. Filos, L. Mendes, G. Serbes, S. Ulukaya, Y. P. Kahya, N. Jakovljevic, T. L. Turukalo, I. M. Vogiatzis, E. Perantoni, *et al.*, "An open access database for the evaluation of respiratory sound classification algorithms," *Physiological measurement*, vol. 40, no. 3, p. 035001, 2019.

[4] R. X. A. Pramono, S. Bowyer, and E. Rodriguez-Villegas, "Automatic adventitious respiratory sound analysis: A systematic review," *PloS one*, vol. 12, no. 5, p. e0177926, 2017.

[5] Y. Ma, X. Xu, and Y. Li, "Lungrn+ nl: An improved adventitious lung sound classification using non-local block resnet neural network with mixup data augmentation." in *Interspeech*, 2020, pp. 2902–2906.

[6] J. Li, J. Yuan, H. Wang, S. Liu, Q. Guo, Y. Ma, Y. Li, L. Zhao, and G. Wang, "LungAttn: Advanced lung sound classification using attention mechanism with dual TQWT and triple STFT spectrogram," *Physiological Measurement*, vol. 42, no. 10, p. 105006, 2021.

[7] Z. Wang and Z. Wang, "A domain transfer based data augmentation method for automated respiratory classification," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 9017–9021.

[8] S. Gairola, F. Tom, N. Kwatra, and M. Jain, "Respirenet: A deep neural network for accurately detecting abnormal lung sounds in limited data setting," in *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, 2021, pp. 527–530.

[9] L. Xu, J. Cheng, J. Liu, H. Kuang, F. Wu, and J. Wang, "Arsc-net: Adventitious respiratory sound classification network using parallel paths with channel-spatial attention," in *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2021, pp. 1125–1130.

[10] T. Nguyen and F. Pernkopf, "Lung sound classification using co-tuning and stochastic normalization," *IEEE Transactions on Biomedical Engineering*, vol. 69, no. 9, pp. 2872–2882, 2022.

[11] I. Moummad and N. Farrugia, "Learning audio features with metadata and contrastive learning," *arXiv preprint arXiv:2210.16192*, 2022.

[12] A. Baade, P. Peng, and D. Harwath, "MAE-AST: Masked Autoencoding Audio Spectrogram Transformer," in *Interspeech*, 2022, pp. 2438–2442.

[13] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Byol for audio: Self-supervised learning for general-purpose audio representation," in *IJCNN*, Jul 2021.

[14] Y. Gong, C.-I. Lai, Y.-A. Chung, and J. Glass, "SSAST: Self-Supervised Audio Spectrogram Transformer," in *AAAI*, vol. 36, no. 10, 2022, pp. 10 699–10 709.

[15] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked modeling duo: Learning representations by encouraging both networks to model the input," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2023, pp. 1–5.

[16] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[17] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*, 2009, pp. 248–255.

[18] J. F. Gemmeke, D. P. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio set: An ontology and human-labeled dataset for audio events," in *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, 2017, pp. 776–780.

[19] Q. Kong, Y. Cao, T. Iqbal, Y. Wang, W. Wang, and M. D. Plumbley, "Panns: Large-scale pretrained audio neural networks for audio pattern recognition," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 28, pp. 2880–2894, 2020.

[20] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio Spectrogram Transformer," in *Interspeech*, 2021, pp. 571–575.

[21] S. Bae, J.-W. Kim, W.-Y. Cho, H. Baek, S. Son, B. Lee, C. Ha, K. Tae, S. Kim, and S.-Y. Yun, "Patch-mix contrastive learning with audio spectrogram transformer on respiratory sound classification," *arXiv preprint arXiv:2305.14032*, 2023.

[22] K. He, X. Chen, S. Xie, Y. Li, P. Dollár, and R. Girshick, "Masked autoencoders are scalable vision learners," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 16 000–16 009.

[23] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar, *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[24] X. Xu and J. Hao, "Multi-layer feature fusion convolution network for audio-visual speech enhancement," *arXiv preprint arXiv:2101.05975*, 2021.

[25] B. Yu, Z. Zhang, D. Zhao, and Y. Wang, "Audio-visual speech enhancement with deep multi-modality fusion," in *2022 5th International Conference on Information Communication and Signal Processing (ICICSP)*. IEEE, 2022, pp. 143–147.

[26] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 3560–3569.

[27] X. Lv, H. Xia, N. Li, X. Li, and R. Lan, "Mfvt: Multilevel feature fusion vision transformer and ramix data augmentation for fine-grained visual categorization," *Electronics*, vol. 11, no. 21, p. 3552, 2022.

[28] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Composing general audio representation by fusing multilayer features of a pre-trained model," in *2022 30th European Signal Processing Conference (EUSIPCO)*, 2022, pp. 200–204.

[29] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[30] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 4700–4708.

[31] A. E. Orhan and X. Pitkow, "Skip connections eliminate singularities," *arXiv preprint arXiv:1701.09175*, 2017.

[32] A. Pasad, J.-C. Chou, and K. Livescu, "Layer-wise analysis of a self-supervised speech representation model," in *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2021, pp. 914–921.

[33] H. Zhang, C. Wu, Z. Zhang, Y. Zhu, H. Lin, Z. Zhang, Y. Sun, T. He, J. Mueller, R. Manmatha, *et al.*, "Resnest: Split-attention networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 2736–2746.

[34] D. S. Park, W. Chan, Y. Zhang, C.-C. Chiu, B. Zoph, E. D. Cubuk, and Q. V. Le, "Specaugment: A simple data augmentation method for automatic speech recognition," *arXiv preprint arXiv:1904.08779*, 2019.