Sound Event Detection using Time-frequency Bounding Boxes with a Self-Supervised Audio Spectrogram Transformer

Zhi Zhu, Yoshinao Sato

Fairy Devices Inc., Japan

***Abstract*—Time-frequency bounding boxes of sound events in audio spectrograms capture essential information across various domains. Nevertheless, previous studies of sound event detection (SED) have focused on detecting temporal boundaries. Although object detection models can be adapted for time-frequency SED, they are intended for image data that exhibit significantly different features compared to audio spectrograms. To address this modality gap, this study employs an audio spectrogram transformer (AST) as the backbone of a detection transformer (DETR). A dataset of whistle sounds from defective wind turbine blades was used for model training and evaluation. The proposed model outperformed a faster region-based convolutional neural network with a residual network backbone, which is an established object detection model. The integration of deformable attention with a multiscale feature pyramid significantly contributed to improving the performance. These results demonstrate the effectiveness of deformable DETR models with an AST backbone for time-frequency SED. Our findings will contribute to advancing time-frequency SED, an area that remains underexplored.**

***Index Terms*—Sound event detection, time-frequency bounding box, audio spectrogram transformer**

1. INTRODUCTION

Sound event detection (SED) aims to identify the time boundaries (i.e., onset and offset times) of specific sound events in audio recordings. However, existing research has largely overlooked an important aspect: the frequency ranges of these events. Time-frequency SED (i.e., detecting the time-frequency bounding boxes of sound events in spectrograms) holds significant value across various domains. For instance, in bioacoustics, a fully annotated bird song dataset includes expert-annotated time-frequency bounding boxes [1], [2]. These boxes serve as acoustic units; understanding how acoustic units are organized into higher-level patterns poses a significant challenge [3]. Moreover, in prognostics and health management, characteristic patterns of sound events can reveal valuable information about machine status. In some cases, temporal SED is insufficient as it discards critical information encoded in the frequency domain. Therefore, accurate detection of time-frequency bounding boxes can assist human investigators and enable in-depth analysis using machine learning. Despite its significance, the detection of time-frequency bounding boxes remains largely unexplored.

Object detection models originally developed for images, such as region-based convolutional neural networks (R-CNN) [4]–[6], the You Only Look Once (YOLO) series, and detection transformers (DETR) [7]–[11] can be repurposed for SED by considering audio spectrograms as images, as observed in a few earlier studies. For instance, Faster R-CNN and YOLO have been applied to detect bioacoustic events [12]–[14] and whistle sounds generated by wind turbine blades [15]. Moreover, a one-dimensional variant of DETR was introduced for detecting time boundaries of sound events [16], [17].

However, using object detection models for SED is challenging due to distinct differences between audio and image modalities. One major difference is the absence of color channels in spectrogram data, unlike image data. Previous studies converted audio spectrograms into RGB images using fixed mappings or learnable pointwise convolution layers, introducing unnecessary arbitrariness and complexity. Another significant issue arises from the feature extraction networks (i.e., backbones). Most object detection models employ pre-trained backbones from large-scale image datasets, such as a residual network (ResNet) [18], which are ill-suited for audio spectrograms because of their unique patterns. Unlike images, the dimensions of spectrograms represent physical properties (time and frequency). Consequently, standard image augmentation methods such as flipping, rotation, scaling, shearing, and translation are not directly applicable to spectrograms.

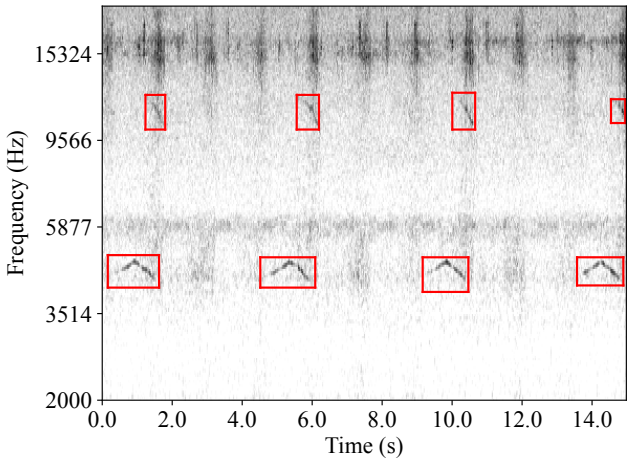

Fig. 1: Example of sound events in spectrogram

To bridge the modality gap, we employ audio spectrogram transformers (AST) [19]–[25] as the backbone of our model. AST models, tailored for audio spectrograms and based on the vision transformer (ViT) architecture, are trained on large-scale audio datasets, such as AudioSet [26]. Hence, they effectively replace conventional backbone models designed for image processing in time-frequency SED. Specifically, we examine a DETR with improved denoising anchor boxes (DINO) [11] and a self-supervised AST model named efficient audio transformer (EAT) [25] for detecting time-frequency bounding boxes in audio spectrograms.

As an example of time-frequency SED, we focus on detecting surface defects on wind turbine blades. Such defects, including cracks and holes, generate sharp whistle-like aerodynamic noises, indicating potential structural issues [15], [27]–[29]. Figure 1 illustrates an example of the sound events to be detected in spectrograms. In-depth information about defects, including their severities, could be estimated from sound event characteristics, such as the duration, frequency center, bandwidth, and peak shape in the spectrograms. Hence, detecting the time-frequency bounding boxes of these whistle sounds, as opposed to merely identifying their time boundaries, is key. The audio signals generated by the wind turbine blades are captured using microphones mounted at the base of the turbine towers.

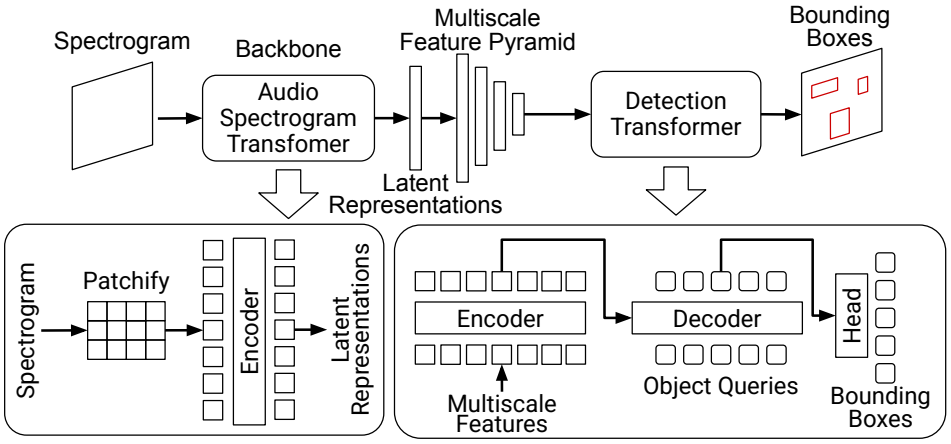

Fig. 2: Structure of proposed model. The left- and right-bottom boxes show the detailed structure of the audio spectrogram transformer (AST) and detection transformer (DETR) models, respectively.

This damage detection method, based on acoustic signals, enables cost-effective, non-destructive, and continuous monitoring of blade conditions.

This study elucidates time-frequency SED, a relatively under-explored area despite its significance. We demonstrate that using an AST model as the backbone for DETR models enhances the detection of sound events in audio spectrograms. The findings in this study could advance time-frequency SED as a fundamental technique across various domains, such as environmental sound recognition, bioacoustics, prognostics, and health management.

2. MODEL

The proposed model integrates an AST backbone with a DETR model, as illustrated in Fig. 2. The model receives a spectrogram as input and outputs the positions and scores of the detected bounding boxes.

For the backbone, we employ EAT, a self-supervised AST model [25]. The AST model partitions the input spectrogram into patches. The encoder extracts the latent audio representations for each patch. During self-supervised training, some patches are randomly masked, and the decoder reconstructs the original spectrogram from the latent representation. Only the pre-trained encoder is used as the backbone of the proposed model without masking. Thus, the outputs of the AST backbone are the latent representations yielded from the final layer of the encoder.

These latent representations are transformed into a multiscale feature pyramid before being input into the DETR model. This method builds on a previous study that applied ViT for object detection [30]. One limitation of ViT and AST is that they produce single-scale latent representations, which hinders their application in this field. However, the multiscale feature pyramid effectively addresses this limitation without tailoring the backbone architecture a priori during pre-training.

We use DINO [11], an extension of DETR, for the detection model. DETR is an end-to-end object detection model with an encoder-decoder structure. The multiscale feature pyramid is fed to the DETR encoder, projected to the same dimension, augmented with positional embeddings, flattened, and concatenated across scales. The decoder, using a fixed learnable object queries, attends to the encoder's output. Finally, a feed-forward network, called the detection head, predicts each object query's position and classification scores. The bounding box positions are specified by the time center, duration, frequency center, and bandwidth. Notably, a "no object" class is included in the object classes. As we do not differentiate between sound event classes, the detection head outputs whether each bounding box represents a sound event. The overall model yields the bounding boxes and scores for the detected sound events.

3. DATA

3.1. Recording

We collected recordings of sounds produced by land-based wind turbines at two locations for 64 days over a year. A waterproof box housing eight micro-electro-mechanical microphones was mounted at the tower's base. Audio signals of 180 to 900 s durations were recorded intermittently at intervals of at least 15 min at a sampling rate of 48 kHz and bit depth of 16. Consequently, 6,026 audio clips totaling 1,147 h were recorded. After the recording, minimum variance distortion-free response beamforming was applied to enhance the sounds generated by the blades.

3.2. Data selection

All recordings were split into 15-second segments. Subsequently, 4,210 segments were selected to balance the following factors: recording date, average spectral flatness over time, standard deviation of loudness over time, and the score for the presence of whistle sounds. The scores were estimated by an audio classification model trained using 240 randomly selected short segments, binary-labeled for the presence or absence of whistle sounds. An attention-based convolutional recurrent neural network [31] was then trained using this small dataset, achieving an area under the precision–recall curve of 0.94 using 5-fold cross-validation. This model was used only to select the data to be annotated.

3.3. Annotation

After data selection, nine reliable crowd workers annotated the spectrograms with time-frequency bounding boxes of whistle sounds. Each short segment was annotated by one worker under the supervision of a single overseer. We compiled 1,843 audio clips, each lasting 15 s, containing 14,420 time-frequency bounding boxes. Notably, we did not distinguish patterns of sound events because the class structure of surface defects on wind turbine blades is unknown. Moreover, the characteristics of whistle sounds are significantly affected by the defect shapes and surrounding environments, including the blade rotation speeds and wind speeds. Therefore, in contrast to conventional object detection tasks, our dataset features only one target class.

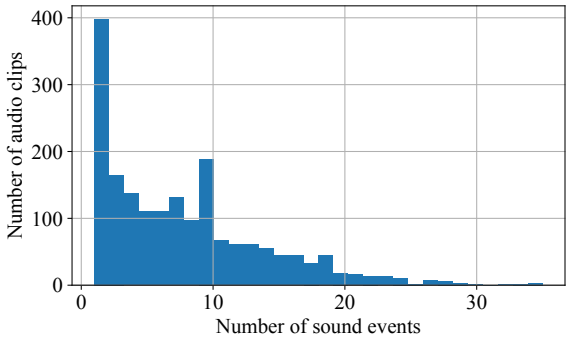

Fig. 3: Distribution of the number of sound events per audio clip

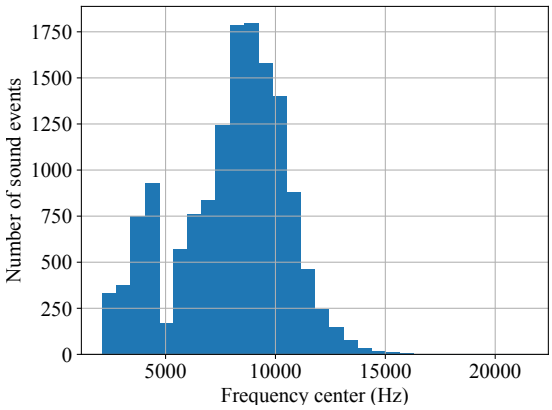

Fig. 4: Distribution of center frequencies

3.4. Analysis

The distribution of the number of annotated bounding boxes per audio clip is shown in Fig. 3. The mean and maximum number of sound events per audio clip are 7.8 and 35, respectively. As described in Section 4.1, we set the number of object queries to 100, which is sufficient for detecting all sound events in a given audio clip in unseen environments. The distribution of the frequency center of the sound events is shown in Fig. 4, with occurrences between 2 and 20 kHz. Therefore, we maintained a sampling rate of 48 kHz in our experiments to ensure that the observed sound events are not lost. The distributions of the duration and bandwidth of the sound events are shown in Fig. 5. As described in Section 4.1, we set the AST patch size to 16 (measured in the number of time-frequency bins) and window shift to 10 ms. Considering this setup, the bounding box width (i.e., the duration of the sound events) is large, and the height (i.e., bandwidth) is small. Hence, low resolution is sufficient in the time direction, but high resolution is crucial in the frequency direction. These characteristics necessitate a multiscale feature pyramid in the model architecture.

4. EXPERIMENTS

We trained and evaluated the proposed DINO with EAT model on wind turbine audio data. A Faster R-CNN [6] with a ResNet [18] backbone served as the baseline. This architecture and its lightweight variants have been applied in prior time-frequency SED studies [12]–[15]. Moreover, an ablation study assessed the effects of dynamic anchor boxes [8], query denoising [9], multiscale feature pyramids [30], and deformable attention [10], with the backbone fixed to EAT.

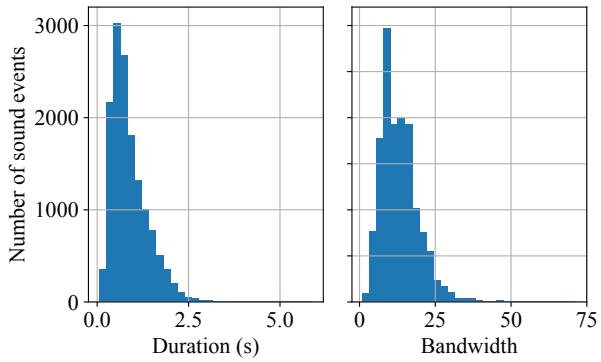

Fig. 5: Distribution of bounding box sizes. Bandwidths are measured in terms of the number of frequency bins.

Table 1: Results of time-frequency SED

Model	Backbone	AP_{50}
Faster R-CNN	ResNet	0.365
Faster R-CNN	EAT	0.370
DINO	ResNet	0.452
DINO	EAT	**0.494**

4.1. Setup

The audio data were split into 15-second segments, converted to log-Mel spectrograms using the short-time Fourier transform with a 25 ms window size and a 10 ms shift, resulting in 182 frequency bins. Unlike in most previous studies on SED, the audio data maintained a 48 kHz sampling rate, yielding spectrogram sizes of $(T, F) = (1498, 182)$.

For the baseline, we employed ResNet-50 with a depth of 50, while the proposed model used an EAT model based on ViT-B/16. A patch size of 16×16 is commonly used in ASTs. The latent representation dimension was $C = 768$, resulting in a shape of $(C, \lfloor T/16 \rfloor, \lfloor F/16 \rfloor) = (768, 93, 11)$. We pre-trained the EAT backbone on the public AudioSet [26] dataset at a 48 kHz sampling rate, rather than using a public model pre-trained at 16 kHz. Moreover, we substituted the one-dimensional sinusoidal positional encoding with two-dimensional encoding. The durations of the audio clips in AudioSet and the wind turbine audio dataset were 10 and 15 s, respectively. The difference in duration was bridged by extending the positional embedding along the time direction during fine-tuning, as described in [23]. The $1/16$ scale latent representations were converted into a multiscale feature pyramid scaled to $\{1/32, 1/16, 1/8, 1/4\}$, following [30]. In DETR, the number and dimension of object queries were set to 100 and 256, respectively, with other setups based on the original work that proposed DINO [11].

The ResNet backbone, Faster R-CNN, and DETRs, except the detection head, were initialized using pre-trained weights from the common objects in context (COCO) dataset [32]. The backbone was not frozen during training. The wind turbine audio data were randomly divided into training, validation, and test sets at a ratio of 7:1:2. The validation set was used to identify the best epoch. For evaluation, we followed the standard evaluation method for the COCO dataset. Specifically, the performance was measured using average precision at an intersection of union (IoU) threshold of 50%, denoted as AP_{50}.

Table 2: Results of ablation study

Model	DAB	DN	MFP + Deformable	AP_{50}
DETR				0.296
DAB-DETR	✓			0.326
DN-DAB-DETR	✓	✓		0.354
Deformable DETR			✓	0.405
DINO	✓	✓	✓	**0.494**

DAB: Dynamic anchor box, DN: Query denoising, MFP: Multiscale feature pyramid, Deformable: Deformable attention.

4.2. Results

In this section, we present the findings of our experiments. The experimental results in Table 1 show that the proposed DINO with EAT model outperformed the baseline models (Faster R-CNN with ResNet). Moreover, the EAT backbone significantly improved the performance of DINO, while the magnitude of improvement was marginal for Faster R-CNN. This result indicates that using AST as a backbone effectively adapts object detection models for audio spectrograms.

The results of the ablation study are summarized in Table 2. According to the experimental results, the deformable attention with a multiscale feature pyramid significantly contributed to the improved performance. Dynamic anchor boxes and query denoising further improved the performance when used with deformable attention. As described in Section 3.4, our data contained small bounding boxes relative to the backbone's latent representation resolution. The multiscale feature pyramid mitigated this issue, consistent with prior research [30]. Notably, deformable attention was proposed to reduce complexity and improve learning convergence associated with multiscale features [10]. Our results affirm that these methods are effective not only for object detection in images but also for time-frequency SED.

4.3. Error analysis

The distribution of the detected bounding box displacements from the actual boxes is shown in Fig. 6. Predictions with scores over 0.5 were counted without applying an IoU threshold. We observed relatively large misplacements compared with typical object detection in images. This result supports the observation that the boundaries of the sound events are blurred and difficult to determine. Hence, setting the IoU threshold to a small value is feasible.

Examples of sound events detected by the human annotators and the proposed model are shown in Fig. 7. Two main errors are observed: First, a single human-annotated box is detected as multiple boxes. Second, multiple human-annotated boxes are detected as one. These misdetections stem from difficulty distinguishing sound event units within complex spectrogram patterns.

5. CONCLUSION

This study elucidates time-frequency SED, which has remained under-explored despite its significance. We adopted object detection models for time-frequency SED, employing an AST backbone to remove the modality gap between image and audio. The experimental results demonstrated that DINO with EAT outperforms the conventional Faster R-CNN with ResNet baseline. The ablation study revealed that the multiscale feature pyramid and deformable attention effectively mitigate resolution issues and enhance performance. These findings indicate the effectiveness of self-supervised ASTs as the backbone of deformable DETRs for time-frequency SED, which can advance research on time-frequency SED in various applications.

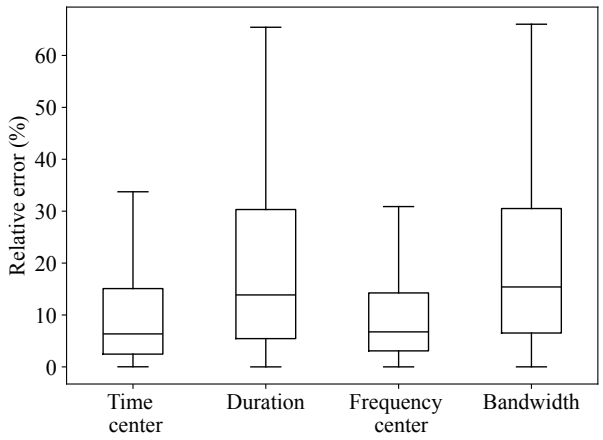

Fig. 6: Distribution of bounding box misalignment relative to the actual duration and bandwidth

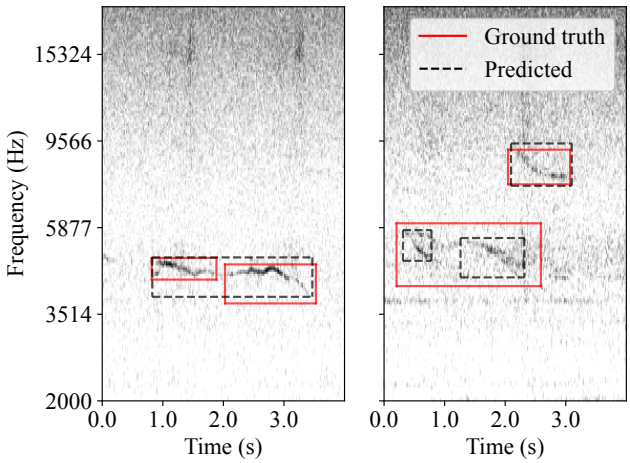

Fig. 7: Examples of detection errors

This research can be extended considering the following aspects. One primary challenge in time-frequency SED is data scarcity. Research on semi-supervised learning to utilize unlabeled and weakly labeled data is a promising direction, as demonstrated in prior studies on object detection in images [33] and temporal SED [17]. Transfer learning is another promising approach. For instance, knowledge learned from the acoustic data of various species, such as birds, frogs, insects, and marine mammals, can be transferred to prognostics and vice versa. Fully annotated bioacoustic datasets, such as BirdCLEF [1] and BirdSet [2], are available for this purpose. Moreover, developing efficient methods for merging and splitting bounding boxes can aid in addressing ambiguities in distinguishing sound event units.

ACKNOWLEDGMENT

The authors thank Yokogawa Electric Corporation for their cooperation in collecting the wind turbine audio data.

REFERENCES

[1] S. Kahl *et al.*, "Overview of BirdCLEF 2019: large-scale bird recognition in soundscapes," in *CLEF*, vol. 2380, no. 256. CEUR, 2019.
[2] L. Rauch *et al.*, "BirdSet: A large-scale dataset for audio classification in avian bioacoustics," in *ICRL*, 2025.

[3] A. Kershenbaum *et al.*, "Acoustic sequences in non-human animals: a tutorial review and prospectus," *Biological Reviews*, vol. 91, no. 1, pp. 13–52, 2016.
[4] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *CVPR*, 2014.
[5] R. Girshick, "Fast R-CNN," in *ICCV*, 2015.
[6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *NeurIPS*, 2015.
[7] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *ECCV*, 2020.
[8] S. Liu, F. Li, H. Zhang, X. Yang, X. Qi, H. Su, J. Zhu, and L. Zhang, "DAB-DETR: Dynamic anchor boxes are better queries for DETR," in *ICLR*, 2022.
[9] F. Li, H. Zhang, S. Liu, J. Guo, L. M. Ni, and L. Zhang, "DN-DETR: Accelerate DETR training by introducing query denoising," in *CVPR*, 2022, pp. 13 619–13 627.
[10] X. Zhu, W. Su, L. Lu, B. Li, X. Wang, and J. Dai, "Deformable DETR: Deformable transformers for end-to-end object detection," in *ICLR*, 2021.
[11] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "DINO: DETR with improved denoising anchor boxes for end-to-end object detection," in *ICLR*, 2023.
[12] K. R. Coffey, R. E. Marx, and J. F. Neumaier, "DeepSqueak: a deep learning-based system for detection and analysis of ultrasonic vocalizations," *Neuropsychopharmacology*, vol. 44, no. 5, pp. 859–868, 2019.
[13] S.-H. Wu, H.-W. Chang, R.-S. Lin, and M.-N. Tuanmu, "SILIC: A cross database framework for automatically extracting robust biodiversity information from soundscape recordings based on object detection and a tiny training dataset," *Ecological Informatics*, vol. 68, p. 101534, 2022.
[14] Q. Hamard, M.-T. Pham, D. Cazau, and K. Heerah, "A deep learning model for detecting and classifying multiple marine mammal species from passive acoustic data," *Ecological Informatics*, vol. 84, p. 102906, 2024.
[15] Y. Zhu and X. Liu, "A lightweight CNN for wind turbine blade defect detection based on spectrograms," *Machines*, vol. 11, no. 99, 2023.
[16] Z. Ye, X. Wang, H. Liu, Y. Qian, R. Tao, L. Yan, and K. Ouchi, "Sound event detection transformer: An event-based end-to-end model for sound event detection," *arXiv e-prints*, 2021.
[17] ——, "SP-SEDT: Self-supervised pre-training for sound event detection transformer," *arXiv e-prints*, 2021.
[18] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016, pp. 770–778.
[19] Y. Gong, Y.-A. Chung, and J. Glass, "AST: Audio spectrogram transformer," in *Interspeech 2021*, 2021, pp. 571–575.
[20] D. Niizumi, D. Takeuchi, Y. Ohishi, N. Harada, and K. Kashino, "Masked spectrogram modeling using masked autoencoders for learning general-purpose audio representation," in *PMLR*, vol. 166, 2022, pp. 1–24.
[21] P.-Y. Huang *et al.*, "Masked autoencoders that listen," in *NeurIPS*, vol. 35, 2022, pp. 28 708–28 720.
[22] 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.
[23] A. Baade, P. Peng, and D. Harwath, "MAE-AST: Masked autoencoding audio spectrogram transformer," in *INTERSPEECH*, 2022, pp. 2438–2442.
[24] D. Chong, H. Wang, P. Zhou, and Q. Zeng, "Masked spectrogram prediction for self-supervised audio pre-training," in *ICASSP*, 2023.
[25] W. Chen, Y. Liang, Z. Ma, Z. Zheng, and X. Chen, "EAT: Self-supervised pre-training with efficient audio transformer," in *IJCAI*, 2024, pp. 3807–3815.
[26] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, and R. C. Moore, "Audio Set: An ontology and human-labeled dataset for audio events," in *ICASSP*, 2017.
[27] Y. Zhang, Y. Cui, Y. Xue, and Y. Liu, "Modeling and measurement study for wind turbine blade trailing edge cracking acoustical detection," *IEEE Access*, vol. 8, pp. 105 094–105 103, 2020.
[28] T.-C. Tsai and C.-N. Wang, "Acoustic-based method for identifying surface damage to wind turbine blades by using a convolutional neural network," *Measurement Science and Technology*, vol. 33, no. 8, p. 085601, 2022.
[29] C. Yang, S. Ding, and G. Zhou, "Wind turbine blade damage detection based on acoustic signals," *Scientific Reports*, vol. 15, no. 1, p. 3930, 2025.
[30] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," in *ECCV*, 2022.
[31] Z. Zhang, S. Xu, S. Zhang, and S. Cao, "Attention based convolutional recurrent neural network for environmental sound classification," *Neurocomputing*, vol. 453, 2021.
[32] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollar, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014.
[33] Z. Dai, B. Cai, Y. Lin, and J. Chen, "UP-DETR: Unsupervised pre-training for object detection with transformers," in *CVPR*, 2021.