applied sciences

Article

Static Sound Event Localization and Detection Using Bipartite Matching Loss for Emergency Monitoring

Chanjun Chun [1], Hyung Jin Park [2] and Myoung Bae Seo [2,*]

[1] Department of Computer Engineering, Chosun University, Gwangju 61452, Republic of Korea; cjchun@chosun.ac.kr

[2] Department of Future & Smart Construction Research, Korea Institute of Civil Engineering and Building Technology (KICT), Goyang-si 10223, Republic of Korea; parkhyungjin@kict.re.kr

* Correspondence: smb@kict.re.kr

Abstract: In this paper, we propose a method for estimating the classes and directions of static audio objects using stereo microphones in a drone environment. Drones are being increasingly used across various fields, with the integration of sensors such as cameras and microphones, broadening their scope of application. Therefore, we suggest a method that attaches stereo microphones to drones for the detection and direction estimation of specific emergency monitoring. Specifically, the proposed neural network is configured to estimate fixed-size audio predictions and employs bipartite matching loss for comparison with actual audio objects. To train the proposed network structure, we built an audio dataset related to speech and drones in an outdoor environment. The proposed technique for identifying and localizing sound events, based on the bipartite matching loss we proposed, works better than those of the other teams in our group.

Keywords: deep learning; sound event localization and detection; convolutional neural network; bipartite matching loss

check for updates

Citation: Chun, C.; Park, H.J.; Seo, M.B. Static Sound Event Localization and Detection Using Bipartite Matching Loss for Emergency Monitoring. *Appl. Sci.* **2024**, *14*, 1539. https://doi.org/10.3390/app14041539

Academic Editor: Edoardo Piana

Received: 2 January 2024
Revised: 8 February 2024
Accepted: 13 February 2024
Published: 14 February 2024

1. Introduction

The advancement of drone technology has revolutionized traditional observation methods. Drones have transcended beyond mere recreational activity to become vital tools in a wide range of applications spanning industries, research, and safety sectors [1]. Specifically, the attachment of advanced sensors to drones has significantly improved the efficiency and effectiveness of surveillance, inspection, and data collection tasks. One of the capabilities that drones have acquired, thanks to advancements in sensor technology, is object detection [2]. Utilizing visual sensors like cameras, drones can identify, track, and analyze specific objects. For instance, in agriculture, drones are used to monitor the health of crops, while in wildlife protection, they can detect poaching activities [3]. In urban environments, their applications vary from analyzing traffic flow and detecting parking violations to even monitoring crowds during large events. However, the scope of drone applications does not stop at visual surveillance. Recently, the capability for sound event detection using auditory sensors like microphones has gained attention. This technology enables drones to collect information based on sound, allowing them to detect specific noises such as abnormal sounds like emergency sirens, gunshots, or screams in urban settings for emergency monitoring [4].

Sound event detection plays a crucial role in transforming drones into monitoring tools for emergency situations [5]. Drones can collect sounds from hazardous areas without direct entry, analyzing them in real time to assess emergencies. Such an analysis can provide valuable information for human rescue, crime response, and disaster management. For example, in the event of natural disasters, drones can be used to capture sounds necessary to find survivors and to identify their locations. This information can help rescue teams allocate resources more effectively and respond swiftly. Moreover, this technology is also useful in noisy environments. In places like airports, railway stations, or large factories where background noise is prevalent, drones can recognize specific noise patterns to distinguish between normal operational sounds and those indicative of emergencies or malfunctions. By integrating sound detection capabilities, drones not only enhance visual assessments but also add an auditory dimension to environmental monitoring, making them even more versatile as tools for ensuring public safety and efficient response to critical situations.

The advent of deep learning significantly revolutionized sound event detection (SED) research [6,7]. The introduction of convolutional neural networks (CNNs) and recurrent neural networks (RNNs) facilitated the processing of complex sound patterns, leading to enhanced detection accuracies [8,9]. This period was also characterized by the development of larger and more diverse datasets, like UrbanSound8K [10] and ESC-50 [11], which allowed for the training of more robust models capable of recognizing a wide range of sound events. Consequently, SED began to find real-world applications in various domains, ranging from smart home systems to urban soundscapes, focusing on detecting specific sound events pertinent to areas like security, healthcare, and environmental monitoring. Recent advancements in SED research have involved the adoption of more complex model architectures, including convolutional recurrent neural networks (CRNNs) and transformers [8,9,12]. These architectures effectively combine the spatial feature extraction capabilities of CNNs with the temporal processing strengths of RNNs or the attention mechanisms of transformers. A significant focus of contemporary research is polyphonic SED, where multiple sound events occur simultaneously [8,9]. Researchers have developed techniques such as multi-label classification and sound source separation to address this challenge. Moreover, the integration of SED with other technologies, such as drones and IoT devices, has broadened its application scope, encompassing areas like emergency response, wildlife monitoring, and industrial inspection.

Sound event localization and detection (SELD) is a task that combines sound event detection (SED) and sound source localization (SSL) [13–17]. SED classifies acoustic events occurring within an audio signal and identifies the start and end points of these classified events. SELD, as a fusion of SED and SSL, not only classifies and detects the activation of sound events within an audio signal but also estimates the location of these activated events. The ability of SELD to discern acoustic events and their locations makes it a valuable tool in scenarios where visual information is absent or inadequate. Therefore, systems equipped with SELD technology are apt for detecting crimes and emergency situations in spaces where individual privacy is paramount. One approach to implementing SELD is to perform SED and SSL independently and then to combine the results from each task. This method can involve a blend of classical algorithms or the application of deep learning models. However, when combining the outcomes of these tasks, a tracking challenge emerges, especially when multiple overlapping events occur simultaneously [18]. This challenge necessitates the correct linkage of the results from each task to their corresponding events. To circumvent this issue, methods that concurrently perform SED and SSL have been developed. These methods can also employ traditional techniques or be based on deep learning algorithms. The advancement and improvement of deep learning algorithms have spurred various research endeavors in the field of deep learning-based SELD, offering potential solutions to previously challenging problems [19].

The end-to-end detection transformer (DETR) represents an innovative paradigm in object detection, predominantly utilized for identifying the positions and classes of objects within images [20]. Distinct from traditional object detection methodologies, DETR circumvents the necessity for anchor boxes and complex post-processing stages [21]. The framework is grounded in the transformer architecture, initially employing a CNN to extract features from images. These features are subsequently fed into the transformer, which concurrently predicts information regarding the object's location and class. This process incorporates the Hungarian algorithm for bipartite matching, culminating in the final prediction [22]. DETR offers a simplistic yet efficient alternative to conventional, intricate object detection systems, demonstrating exceptional performance, particularly in scenarios involving large-scale objects or scenes.

In this paper, we present a novel neural network architecture for static sound event localization and detection, inspired by the principles of detection transformer (DETR) applied in the auditory domain. Specifically, our method is the use of the ResNet-50 model within a CNN framework [23], tailored to process audio signals segmented into 30-second intervals. Unlike traditional object detection methods in computer vision, which rely on numerous grids and anchors, leading to redundant detections refined through techniques like non-maximum suppression (NMS), our approach is anchor- and NMS-free. It infers a fixed-size set of audio predictions, exceeding the actual count of audio objects in the mel-spectrogram, ensuring all audio objects are captured without relying on image-based post-processing methods. This approach is further augmented by bipartite matching using the Hungarian algorithm, aligning the order of audio predictions with the actual audio objects and efficiently handling the classification, localization, and detection (onset and offset) of audio events. This innovative methodology provides a bespoke solution for sound event localization and detection, leveraging advancements in object detection and adapting them creatively to the auditory context.

The structure of our paper is organized as follows: Following the introduction, Section 2 delves into the methodology of sound event localization and detection based on bipartite matching loss. In Section 3, we discuss the experiments conducted to evaluate the performance of our proposed method. Finally, Section 4 concludes the paper, summarizing our findings and contributions to the field of sound event localization and detection.

2. Proposed Static Sound Event Localization and Detection Using Bipartite Matching Loss

The proposed neural network architecture in this paper for static sound event localization and detection is illustrated in Figure 1. Specifically, it utilizes a stereo audio signal with a sampling rate of 48 kHz, which is converted into a mel-spectrogram by performing a 2048-point short-time Fourier transform (STFT), with the hop length and mel-filterbanks meticulously set to 512 and 128, respectively. The architecture fundamentally adopts a convolutional neural network (CNN) framework, within which the ResNet-50 model is actively utilized [23]. Audio signals are meticulously segmented into 30 s increments. Within this framework, a single audio signal may contain only one distinct sound event, or it may not contain any at all. Additionally, there may be instances where multiple distinct sound events are present, and while these events may overlap, we have scrupulously constructed the audio samples to ensure that no more than two sound events overlap simultaneously. Furthermore, the audio signals have been designed with the capacity to contain up to a maximum of nine distinct sound events.

In object detection tasks commonly employed within computer vision, the number of grids and anchors typically results in an abundance of redundant detections, which are then refined to unique results through techniques such as non-maximum suppression (NMS). However, such techniques are tailored to image information like intersection-over-union (IoU) and may not be as effective for audio information, such as mel-spectrograms. Therefore, to adopt anchor- and NMS-free approaches, the neural network is designed to infer the detection of an exact number of N audio predictions. This circumvents the dependency on traditional image-based post-processing methods, presenting a tailored solution for the auditory domain. Here, N is typically set to a number greater than the actual count of audio objects present in the mel-spectrogram. If N is less than the number of audio objects, it becomes challenging to infer all audio objects within that particular audio signal accurately. By choosing a sufficiently large N, it is implied that when there are fewer audio objects than the number set for N, the remaining predictions can be considered as padding with $\varnothing$ (no objects). This approach ensures that the network is always primed to detect up to the maximum expected number of audio objects without the risk of missing any due to an underestimation of N.

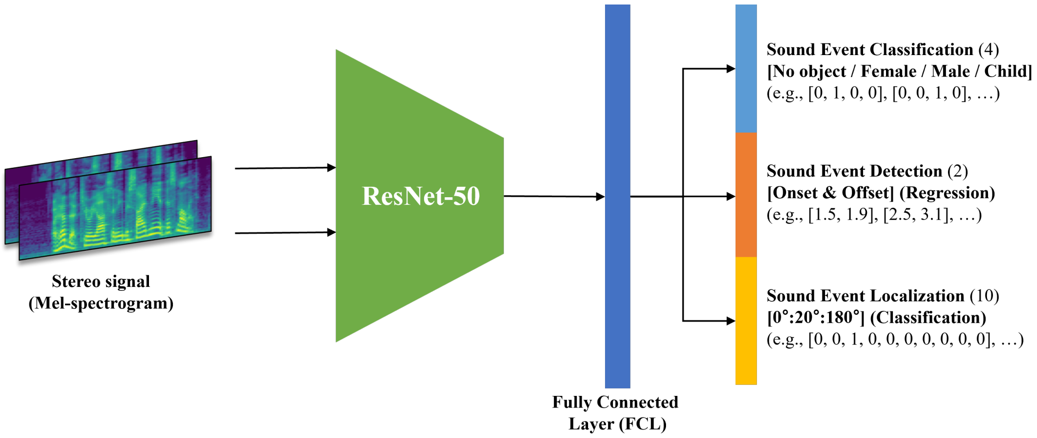

Stereo signal
(Mel-spectrogram)
ResNet-50
Fully Connected Layer (FCL)
Sound Event Classification (4)
[No object / Female / Male / Child]
(e.g., [0, 1, 0, 0], [0, 0, 1, 0], …)
Sound Event Detection (2)
[Onset & Offset] (Regression)
(e.g., [1.5, 1.9], [2.5, 3.1], …)
Sound Event Localization (10)
[0°:20°:180°] (Classification)
(e.g., [0, 0, 1, 0, 0, 0, 0, 0, 0, 0], …)

Figure 1. Overview of proposed sound event localization and detection method using bipartite matching loss. Our method uses the ResNet-50 model within a CNN framework and infers a fixed-size set of audio predictions. This method is enhanced by employing bipartite matching through the Hungarian algorithm, which aligns audio predictions with their corresponding actual audio objects. It effectively manages classification, horizontal localization, and the detection of the beginning and end of audio events.

If N is sufficiently large, $\hat{y}$ will infer a fixed-size set of N results, some of which may be padded with $\varnothing$ (no objects) to fill the quota. To align the order of the inferred N audio predictions with the actual audio objects, y, the bipartite matching is conducted. This process involves finding a permutation of the N elements that result in the lowest loss. The equation for this matching process is formalized as follows:

$$\hat{\sigma} = \arg\min_{\sigma} \sum_{i=1}^{N} \mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)}), \tag{1}$$

where $\mathcal{L}_{\text{match}}(y_i, \hat{y}_{\sigma(i)})$ is a pair-wise matching cost between the actual audio objects and the predictions with index $\sigma(i)$. This optimal assignment can be computed efficiently with the Hungarian algorithm [22].

In calculating the matching cost, three components can be considered: the classification prediction of the audio object, the inference of localization, and the inference results for detection (onset and offset). For the classification of audio objects, four classes are considered: male, female, child, and $\varnothing$ (no object). Localization involves estimating the azimuthal angle on a horizontal plane, discretized into ten classes representing directions at 20-degree intervals from 0 to 180 degrees. For detection, onset and offset information is estimated in a regression manner akin to bounding box estimations in the object detection task. The composite loss equation that incorporates these elements might be represented as follows:

$$\mathcal{L}_{\text{hungarian}}(y, \hat{y}) = \sum_{i=1}^{N} \Big[-\log \hat{p}_{\hat{\sigma}(i)}(c_i) + \mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{\text{loc}}(l_i, \hat{l}_{\hat{\sigma}(i)}) + \mathbb{1}_{\{c_i \neq \varnothing\}} \mathcal{L}_{\text{det}}(d_i, \hat{d}_{\hat{\sigma}(i)}) \Big], \tag{2}$$

where $\hat{\sigma}$ is the optimal assignment computed from the Hungarian algorithm. Here, when c_i corresponds to $\varnothing$ (no objects), it is common practice to assign a smaller weight to such instances to avoid the model biasing towards predicting $\varnothing$ (no objects). In DETR [20], they may assign a weight as low as one-tenth for $\varnothing$ (no objects); however, this paper opts for a relatively larger weight, specifically half that of the other classes. This adjustment is due to the DETR model typically setting N to around 100, which results in a significant number of $\varnothing$ (no objects) predictions. Contrarily, by setting N to 15, this paper limits the number of $\varnothing$ (no objects) instances and therefore experimentally employs a higher weight for these instances. The justification for this approach is based on the reduced relative frequency of

∅ (no objects) cases, allowing for a slight increase in weight without overwhelming the learning process with an excessive focus on the ∅ (no objects) class.

The overall architecture is surprisingly simple and illustrated in Figure 1. In the proposed architecture, a straightforward feed-forward network (FFN) is appended to the fundamental structure of a ResNet-50. This enhancement is designed to process and refine the features extracted by the ResNet-50 for the task at hand, enabling a more sophisticated interpretation of the data pertinent to the objectives of the neural network, specifically for static sound event localization and detection.

For the training of the neural network, the Adam optimizer was employed with the learning rate set to 0.001 [24]. Upon observing that the validation loss stabilized and plateaued, we manually reduced the learning rate by a factor of 10 and resumed the training process. The batch size was configured to 64, and within the CNN architecture, a dropout rate of approximately 0.1 was utilized to facilitate training [20,22]. This approach is consistent with regularization strategies that prevent overfitting and encourage the model to learn more robust features.

3. Experiments

We constructed an environment and collected the dataset for identifying audio signals (male/female/child/no object) using the stereo microphone mounted on a drone. The directional stereo microphone was utilized, and the data acquisition environment is as illustrated in Figure 2. Ideally, attaching the microphone directly to the drone would have been more representative of a real-world scenario; however, due to physical constraints, we opted to place the microphone on a tripod, as shown in the figure, with a drone hovering above them. The drone hovered at an altitude ranging between approximately 2.5 m and 3.0 m, while the tripod was set up at a height of 1.7 m to position the microphone. The loudspeaker was installed at distances of 5 m, 10 m, and 20 m from the microphone, and considering that the speech signals might originate from a position lower than the microphone, the loudspeaker was placed on the ground. Here, we utilized a SONY PCM-A10 voice recorder as a microphone and a GENELEC 8010A as a loudspeaker. The model proposed in this paper also predicts the horizontal angle; hence, it was possible to position the loudspeaker from 0 degrees to 180 degrees at 20-degree intervals. Figure 3a displays the configuration of the loudspeaker, while Figure 3b shows an actual photograph of the recording site. Since rotating the microphone was much more feasible than changing the physical location of the loudspeaker, the recordings were captured by adjusting the distance of the loudspeaker and rotating the microphone to cover the required angles.

The recordings of speech and drone noise were carried out separately, and the collected audio samples were mixed. Although 25 different models of drones were utilized, only about 20 models were actually used in the creation of the dataset. The voice actors for men, women, and children each comprised about 30 individuals, with children being defined as those under the age of 13. This is because students attend elementary school until the age of 13 in Korea. In addition, middle and high school students were not included as their voices contain elements of both children and adults. Each speaker had a set of 20 sentences related to distress situations (e.g., “Help me!”). These recorded voices were appropriately edited and used, with recordings of individual voices created without overlap and any overlapping instances were mixed randomly. Uncontrollable noises captured during outdoor recordings, such as passing vehicles, were excluded from the mix. Only recordings captured in as quiet an environment as possible were utilized, and after mixing various voices, drone noises were also mixed in. Ultimately, a total of 50,000 audio samples each lasting 30 s were created, with 500 samples set aside as a test dataset.

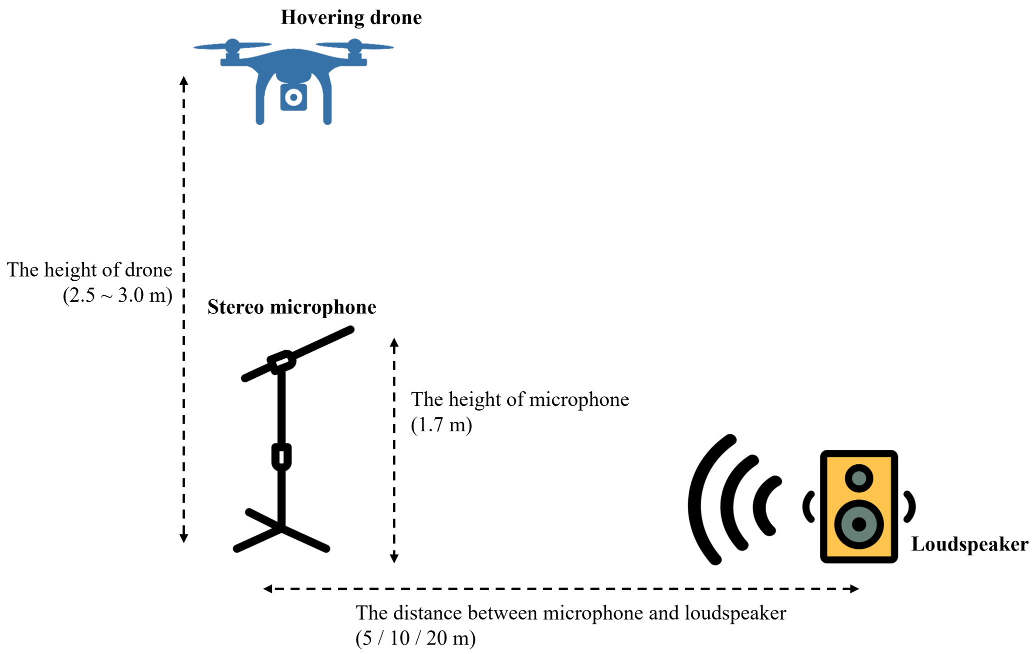
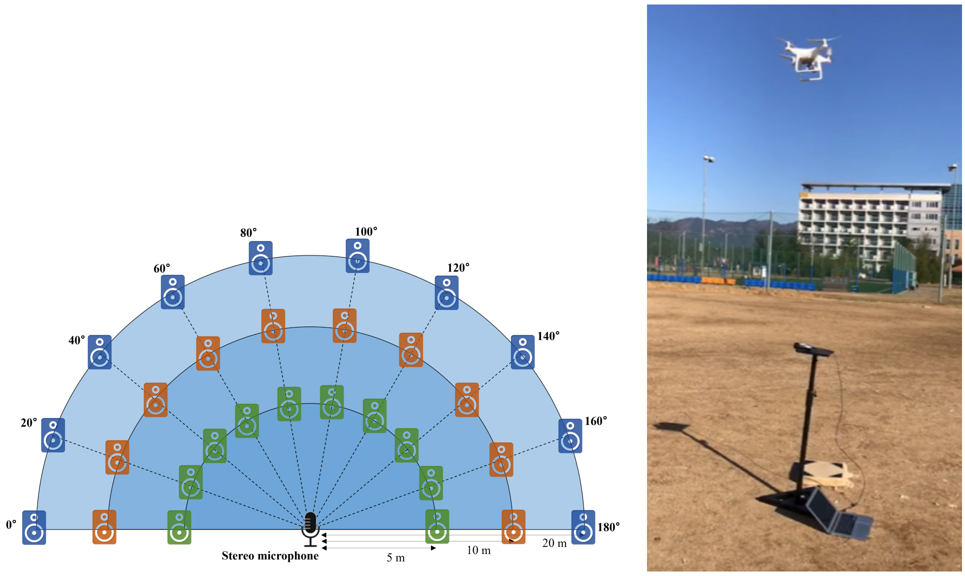


Figure 2. The data acquisition configuration in an outdoor environment. Due to limitations in the physical setup, we chose to mount the microphone on a tripod with a drone flying overhead. The drone maintained a height varying between roughly 2.5 m and 3.0 m, whereas the tripod was adjusted to a height of 1.7 m to properly place the microphone. The loudspeaker was positioned at varying distances from the microphone, specifically at 5, 10, and 20 m, respectively.



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

Figure 3. Recording configuration depending on speaker placement. (**a**) The loudspeaker is positioned at intervals of 20 degrees, ranging from 0 to 180 degrees. (**b**) An actual photograph of the recording site is displayed.

To evaluate the performance of our trained neural network model, we have established a new metric. This proposed metric allows us to assess information about the class and direction of the sound source simultaneously. First, to measure the performance of the estimated direction of origin, we calculate the mean-squared-error (MSE) loss for the directional vector. Here, if the estimated direction is not precise but predicts a neighboring direction, we compensate for this by using a weighted moving average (WMA). The metric for estimating the direction of origin is as follows:

$$WMA(d_j) = 0.05 \cdot d_{j-2} + 0.1 \cdot d_{j-1} + 0.7 \cdot d_j + 0.1 \cdot d_{j+1} + 0.05 \cdot d_{j+2}, \quad j \in [0,9] \tag{3}$$

Based on the compensation described in Equation (3), the performance regarding the direction of origin is measured as follows:

$$D_{\text{direction}} = \sum_{j=0}^{9} (WMA(\hat{d}_j) - WMA(d_j))^2 \tag{4}$$

Here, note that our method involves estimating a total of ten azimuth angles at 20-degree intervals, ranging from 0 to 180 degrees. Consequently, we have ten distinct azimuth angles, labeled from 0 to 9. Afterward, to estimate the performance of sound localization, we structured the source distinction vector analogously to the directional vector. Likewise, we employ the mean-squared-error (MSE) loss to evaluate the efficacy of the sound localization. The corresponding equation is presented below.

$$D_{\text{class}} = \sum_{j=0}^{2} (\hat{c}_j - c_j)^2 \tag{5}$$

Here, the sound sources are categorized into three types: male, female, and child. Thus, the classification range is from 0 to 2. Ultimately, the performance metrics for sound localization and classification are aggregated with weights of 0.8 and 0.2, respectively, to derive the final performance indicator. These weights have been internally determined during the course of the project, such that

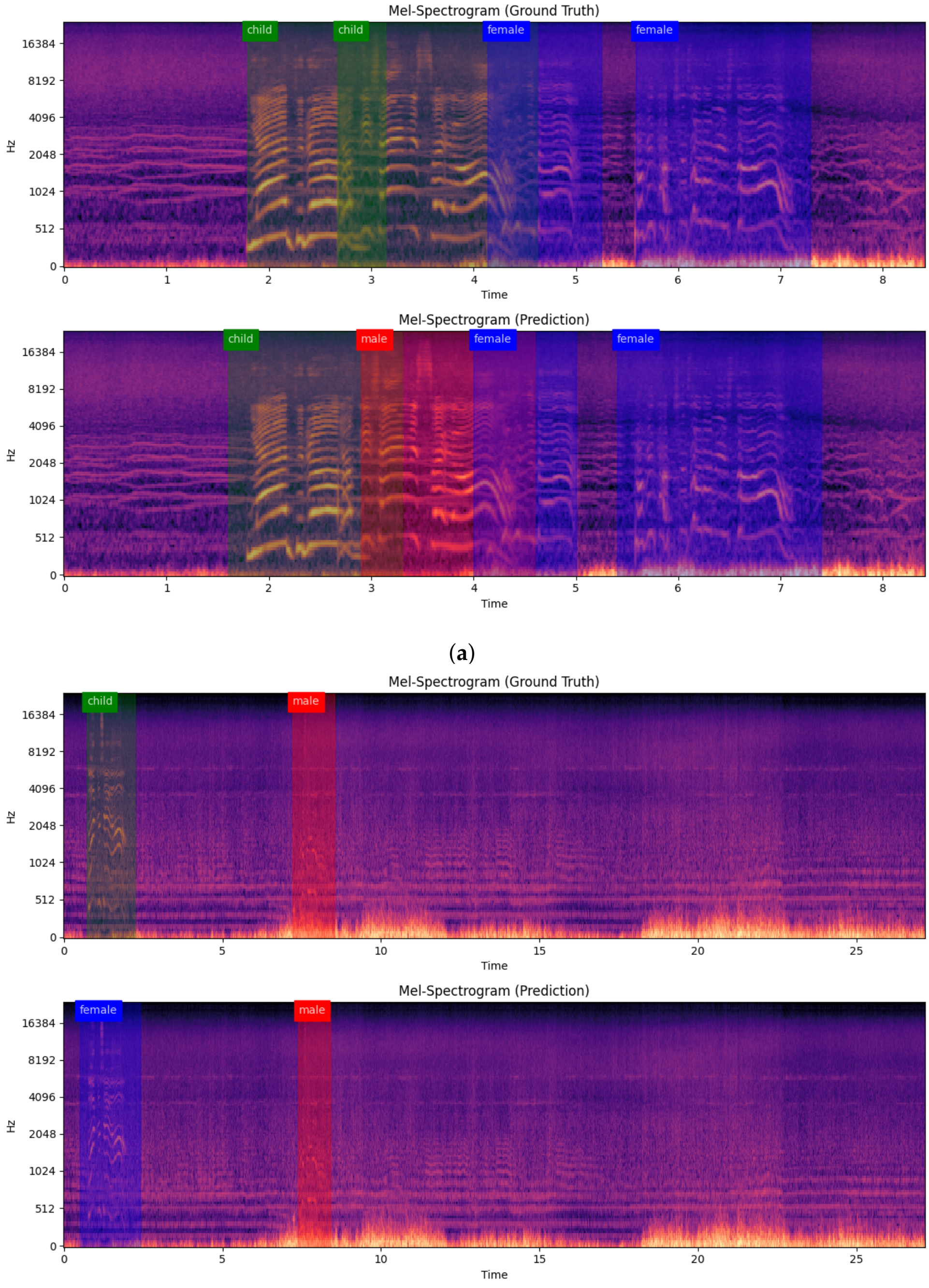
$$D_{\text{total}} = 0.8 \cdot D_{\text{direction}} + 0.2 \cdot D_{\text{class}} \tag{6}$$

Based on the values in Equation (6), a model can be considered to perform better as the number decreases. Table 1 presents the internal team rankings predicated on Equation (6). It has been substantiated that the sound event localization and detection technique, predicated on the proposed bipartite matching loss, exhibits a superior performance relative to that of the other contending teams. Regrettably, an in-depth exploration into the methodologies employed by the competing teams for solving this problem lies beyond the scope of this paper. Nevertheless, an objective assessment based on quantifiable metrics has corroborated the commendable efficacy of the proposed approach. It is pertinent to note that the evaluation dataset was meticulously constructed by teams not affiliated with this internal competition.

Table 1. The internal team rankings predicated on Equation (6).

Teams	D_{total}
Internal Team #1	3.75385
Internal Team #2	3.74085
Internal Team #3	3.69416
Internal Team #4	3.40980
Internal Team #5 (Ours)	3.15351

Figure 4 shows the results of performing SED using the proposed method in this paper. Figure 4a illustrates the ground truth, and Figure 4b shows the predicted results (only left channel), respectively. It is confirmed that the segments where speech is present are detected with high accuracy. In terms of classification, the performance is moderately accurate for males and females, but there were instances of confusion between males or females for children. This is because the voice of a child can exhibit characteristics very similar to those of males or females. Table 2 implies the confusion matrix for classification performance. (a) represents the proposed method, while (b) refers to the results obtained from GoogLeNet [25]. As shown in the table, there are instances where children are misclassified as male or female. It is observable that the proposed method demonstrates relatively better performance compared to GoogLeNet.



(**a**)

(**b**)

Figure 4. The results of performing SED using the proposed method. (**a**) illustrates the ground truth, and (**b**) shows the predicted results (only left channel).

Table 2. Confusion matrix for the classification performance. (**a**) Confusion matrix for the classification performance of the proposed SED method. (**b**) Confusion matrix for the classification performance using GoogLeNet [25].

	(a)		
Pred. / **Actual**	**Child**	**Male**	**Female**
child	4023	127	251
male	98	3949	27
female	87	13	4028
	(b)		
child	3914	194	293
male	118	3930	26
female	162	16	3950

4. Conclusions

This paper introduced an innovative neural network architecture for static sound event localization and detection, adapting the detection transformer (DETR) methodology for auditory analysis. Our approach involves a novel application of the ResNet-50 model within a CNN framework, processing stereo audio signals into mel-spectrograms. By forgoing traditional anchor and non-maximum suppression methods common in visual object detection, our system efficiently infers a set number of audio predictions. This method, enhanced by bipartite matching and the Hungarian algorithm, ensures accurate classification, localization, and detection of sound events. Our experimental findings, based on a dataset created with diverse audio samples including drone noises and human speech, validate the proposed model's effectiveness in sound event localization and detection. The research presents a significant contribution to the field, demonstrating a unique adaptation of DETR principles to the auditory domain. The implications of this study extend to various practical applications, notably in emergency response and environmental monitoring, highlighting the model's potential in real-world scenarios.

In the future, we will develop the proposed method to robustly detect and localize not only static sound events but also dynamically moving sound events and conduct in-depth comparative analysis with various SELD methods. It is implied that our method can be utilized in systems capable of making appropriate responses using microphone signals acquired from drones in disaster or hazardous environments.

Author Contributions: Conceptualization, C.C.; data curation, H.J.P.; funding acquisition, M.B.S.; investigation, H.J.P.; methodology, C.C.; project administration, M.B.S.; software, C.C.; supervision, M.B.S.; validation, H.J.P. and C.C.; visualization, C.C.; writing—original draft preparation, C.C.; writing—review and editing, M.B.S. All authors have read and agreed to the published version of the manuscript.

Funding: This work is supported by the Korea Agency for Infrastructure Technology Advancement (KAIA) grant funded by the Ministry of Land, Infrastructure and Transport (grant 20UUTI-B157786-02).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data is contained within the article.

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

References

1. Hoang, M.L. Smart Drone Surveillance System Based on AI and on IoT Communication in Case of Intrusion and Fire Accident. *Drones* **2023**, *7*, 694. [CrossRef]
2. Huangfu, Z.; Li, S. Lightweight You Only Look Once v8: An Upgraded You Only Look Once v8 Algorithm for Small Object Identification in Unmanned Aerial Vehicle Images. *Appl. Sci.* **2023**, *13*, 12369. [CrossRef]
3. Shah, S.A.; Lakho, G.M.; Keerio, H.A.; Sattar, M.N.; Hussain, G.; Mehdi, M.; Vistro, R.B.; Mahmoud, E.A.; Elansary, H.O. Application of Drone Surveillance for Advance Agriculture Monitoring by Android Application Using Convolution Neural Network. *Agronomy* **2023**, *13*, 1764. [CrossRef]
4. Serrenho, F.G.; Apolinário, J.A.; Ramos, A.L.L.; Fernandes, R.P. Gunshot Airborne Surveillance with Rotary Wing UAV-Embedded Microphone Array. *Sensors* **2019**, *19*, 4271. [CrossRef] [PubMed]
5. Villegas-Ch, W.; Govea, J. Application of Deep Learning in the Early Detection of Emergency Situations and Security Monitoring in Public Spaces. *Appl. Syst. Innov.* **2023**, *6*, 90. [CrossRef]
6. Neri, M.; Battisti, F.; Neri, A.; Carli, M. Sound Event Detection for Human Safety and Security in Noisy Environments. *IEEE Access* **2022**, *10*, 134230–134240. [CrossRef]
7. Mesaros, A.; Heittola, T.; Virtanen, T.; Plumbley, M.D. Sound event detection: A tutorial. *IEEE Signal Process. Mag.* **2021**, *38*, 67–83. [CrossRef]
8. Li, Y.; Liu, M.; Drossos, K.; Virtanen, T. Sound Event Detection Via Dilated Convolutional Recurrent Neural Networks. In Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Barcelona, Spain, 4–8 May 2020; pp. 286–290. [CrossRef]
9. Çakir, E.; Parascandolo, G.; Heittola, T.; Huttunen, H.; Virtanen, T. Convolutional Recurrent Neural Networks for Polyphonic Sound Event Detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2017**, *25*, 1291–1303. [CrossRef]
10. Salamon, J.; Jacoby, C.; Bello, J.P. A Dataset and Taxonomy for Urban Sound Research. In Proceedings of the 22nd ACM International Conference on Multimedia, Orlando, FL, USA, 3–7 November 2014; pp. 1041–1044. [CrossRef]
11. 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. [CrossRef]
12. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS2017), LongBeach, CA, USA, 4–9 December 2017; Volume 30.
13. Nguyen, T.N.T.; Watcharasupat, K.N.; Nguyen, N.K.; Jones, D.L.; Gan, W.S. Salsa: Spatial cue-augmented log-spectrogram features for polyphonic sound event localization and detection. *IEEE/ACM Trans. Audio Speech Lang. Process.* **2022**, *30*, 1749–1762. [CrossRef]
14. Nguyen, T.N.T.; Jones, D.L.; Watcharasupat, K.N.; Phan, H.; Gan, W.S. SALSA-Lite: A fast and effective feature for polyphonic sound event localization and detection with microphone arrays. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 716–720.
15. Shimada, K.; Koyama, Y.; Takahashi, N.; Takahashi, S.; Mitsufuji, Y. ACCDOA: Activity-coupled cartesian direction of arrival representation for sound event localization and detection. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Toronto, ON, Canada, 6–11 June 2021; pp. 915–919.
16. Shimada, K.; Koyama, Y.; Takahashi, S.; Takahashi, N.; Tsunoo, E.; Mitsufuji, Y. Multi-ACCDOA: Localizing and Detecting Overlapping Sounds from the Same Class with Auxiliary Duplicating Permutation Invariant Training. In Proceedings of the International Conference on Acoustics, Speech and Signal Processing (ICASSP), Singapore, 23–27 May 2022; pp. 316–320.
17. Adavanne, S.; Politis, A.; Nikunen, J.; Virtanen, T. Sound event localization and detection of overlapping sources using convolutional recurrent neural networks. *IEEE J. Sel. Top. Signal Process.* **2018**, *13*, 34–48. [CrossRef]
18. Cao, Y.; Kong, Q.; Iqbal, T.; An, F.; Wang, W.; Plumbley, M.D. Polyphonic Sound Event Detection and Localization Using a Two-Stage Strategy. In Proceedings of the Detection and Classification of Acoustic Scenes and Events Workshop (DCASE), New York, NY, USA, 25–26 October 2019; pp. 30–34.
19. Shin, Y.; Kim, Y.G.; Choi, C.H.; Kim, D.J.; Chun, C. SELD U-Net: Joint Optimization of Sound Event Localization and Detection With Noise Reduction. *IEEE Access* **2023**, *11*, 105379–105393. [CrossRef]
20. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-End Object Detection with Transformers. *arXiv* **2020**, arXiv:2005.12872.
21. Hosang, J.; Benenson, R.; Schiele, B. Learning Non-maximum Suppression. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6469–6477. [CrossRef]
22. Kuhn, H.W. The Hungarian method for the assignment problem. *Nav. Res. Logist. Q.* **1955**, *2*, 83–97. [CrossRef]
23. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep Residual Learning for Image Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778. [CrossRef]

24. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv* **2017**, arXiv:1412.6980.
25. Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1–9. [CrossRef]

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.