Enemy Spotted: In-game Gun Sound Dataset for Gunshot Classification and Localization

Junwoo Park, Youngwoo Cho, Gyuhyeon Sim, Hojoon Lee, Jaegul Choo
Kim Jaechul Graduate School of AI
Korea Advanced Institute of Science and Technology (KAIST)
Daejeon, South Korea
{Junwoo.park, cyw314, ghsim, joonleesky, jchoo}@kaist.ac.kr

***Abstract*—Recently, deep learning-based methods have drawn huge attention due to their simple yet high performance without domain knowledge in sound classification and localization tasks. However, a lack of gun sounds in existing datasets has been a major obstacle to implementing a support system to spot criminals from their gunshots by leveraging deep learning models. Since the occurrence of gunshot is rare and unpredictable, it is impractical to collect gun sounds in the real world. As an alternative, gun sounds can be obtained from an FPS game that is designed to mimic real-world warfare. The recent FPS game offers a realistic environment where we can safely collect gunshot data while simulating even dangerous situations. By exploiting the advantage of the game environment, we construct a gunshot dataset, namely BGG, for the firearm classification and gunshot localization tasks. The BGG dataset consists of 37 different types of firearms, distances, and directions between the sound source and a receiver. We carefully verify that the in-game gunshot data has sufficient information to identify the location and type of gunshots by training several sound classification and localization baselines on the BGG dataset. Afterward, we demonstrate that the accuracy of real-world firearm classification and localization tasks can be enhanced by utilizing the BGG dataset.**

***Index Terms*—shooting game, gunshot dataset, gunshot classification and localization**

I. Introduction

The first-person shooter (FPS) game, designed to mimic real-world warfare and combat situations, is a popular game genre. To defeat enemies in the FPS game, a player under attack must decide whether to strike back or retreat by considering the enemies' position and firearms. However, human vision may not be able to capture where the enemies are and what firearms they have as the distance between the player and them increases. In addition, when the enemies camouflage themselves, it is also difficult to detect the enemies by human vision. In these cases, gunshots can be clues for estimating the enemies' state. For example, an expert FPS gamer can recognize the tiny difference in stereophonic sound from a headphone, and she can roughly guess the position and the firearms of the enemies. The reason the player can establish strategy based on auditory information is that a game engine can reproduce the characteristics of the sound that varies with distance and direction. Inspired by the realism of the game, we hypothesize that a prediction model, which localizes enemies and identifies firearms from in-game gunshots, can also be applied to real-world gunshots. Specifically, a support system that can spot the enemies and determine the type of firearm from gunshots does not only assist the beginners of the game, but also aids soldiers and police officers who track the criminals in the real world.

The problem of spotting enemies and identifying the firearms can be formulated as a sound localization and classification task. The sound localization task aims at estimating the distance and angle between a sound source and a receiver, and the sound classification task aims at categorizing the sound into predefined classes. Recently, deep learning has become the standard practice in sound localization [1], [2] and classification [3], [4] tasks. However, a lack of gunshot samples in existing datasets [5], [6] has been a major obstacle to building the support system in surveillance and military service by leveraging the deep learning models. A real-world gunshot dataset, `Forensic` [7], alleviates this problem by firing and recording in a controlled environment. Nevertheless, collecting data in the real world is expensive and may not cover the engagement situation due to the dangers of firearms.

As an alternative, the FPS games enable us to easily collect data by simulating engagement situations in which a bullet passes by the side of the face. By exploiting this advantage of the simulation, we constructed a battleground gunshot (`BGG`) dataset collected from PlayerUnknown's Battleground[1] (PUBG). PUBG is a representative FPS game where 100 players aim to survive by eliminating each other with firearms and other weapons until only one player (or team) remains. We recorded the gun sounds and labeled the types of firearms, their distance, and angle by varying the enemies' position.

Before applying the in-game gunshots to real-world tasks, we evaluated baselines widely used in sound classification and Transformer [8], which is underexplored in gunshot classification and localization tasks. As a result, we confirmed that the in-game gunshots sufficiently reflect the characteristics of the real-world sound and the models can learn those characteristics to predict the location enemies and their firearms. Furthermore, we empirically demonstrated that our in-game gunshot dataset can improve the accuracy of both sound classification and gunshot localization in real-world datasets.

The contributions of our work are summarized as follows:

[1]https://krafton.com/games/battlegrounds

- We construct a gunshot dataset, BGG, for gunshot classification and localization by exploiting the advantages of the game environment.
- We verified that the deep learning models can spot enemies position and identify their firearms from the in-game gunshots by presenting the application deployed in PUBG.
- We demonstrated that the BGG dataset can be utilized to improve the accuracy of the real-world sound classification and gunshot localization tasks.

II. Related work

In this work, we deal with two problems, sound classification and localization, which are the tasks of identifying the type of sound and estimating the location of sound sources. For decades, both tasks have been studied as a crucial component for various domains such as robotics [9], [10], speech recognition [11], and surveillance system [12]. Generally, sound localization is more challenging than sound classification since the number of sources, the interaction between sources and surroundings, and the movement of sources should be considered [13] in the case of localization. The fundamental idea of sound localization is to leverage the difference in arrival time or phase of signals obtained from the microphone array. Currently, well-known sound source localization methods include multiple signal classification (MUSIC) [14] and generalized cross-correlation with phase transform (GCC-PHAT) [15]. Recently, data-driven approaches, (*i.e.,* sound classification and localization based on deep learning), have been actively studied [13]. Most of the deep learning-based approaches employ deep convolutional neural networks (DCNNs) [10] or convolutional recurrent neural networks (CRNNs) [4], and each method varies in the architectural details, the input features, and the output type.

An application of sound classification and localization is gunshot classification and localization, which focuses on gunshots as sound sources, and several systems for detecting gunshots have been proposed for achieving social safety [16], [17], [18]. A major obstacle in the gunshot classification and localization task is the data acquisition problem. Even though there are several datasets containing gunshot samples, the number of gunshot-related audio samples is fewer than that of the other classes. For instance, the number of gun sound samples in AudioSet [5] and UrbanSound8k [6] account for only 0.2% and 4.2%, respectively, since the occurrence of gunshots is rare and unpredictable. Therefore, predictions from deep learning models trained on the existing datasets are inaccurate for gunshots compared to other types of sound. Singh et al. [19] reported that sound source localization models could not distinguish gunshots and gunshot-like audio events (*e.g.,* plastic bag bursting) on the UrbanSound8k. To address this issue, Raponi et al. [7] proposed a gunshot-dedicated dataset, Forensic, a collection of gunshots recorded in a controlled environment. However, the application of the Forensic is limited to a few specific scenarios because the Forensic does not reflect the actual engagement situation. As an alternative, several data-efficient methods were proposed to deal with the class imbalance and data deficiency problems [20], [21]. Shimada et al. [21] introduced a few-shot sound classification method that classifies rare classes (*e.g.,* gunshot, glassbreak, and babycry) by utilizing a metric-learning algorithm.

TABLE I
Basic statistics of BGG dataset. The values in the parentheses are the number of samples and classes that are not gunshots.

	BGG	Foren	Urban
# of Gun sound samples	2,195 (57)	2,241	374 (8,358)
Range of audio length (sec.)	[3,8]	[1.5,2.5]	[0.05,4]
Sampling rate of audio (KHz)	44.1	96	[8,192]
# of Gun categories	37 (1)	18	1 (9)
# of Directions from source	5 (1)	13	-
# of Distances from source	6 (1)	8	-
Range of distance (m)	[0,600]	[0,150]	-

III. BGG dataset

Unlike other types of sound, it is difficult to collect gunshots in real world since firing a gun is a rare and dangerous event. One may collect gunshots in a shooting range, but it is expensive, and the sounds could be reproduced in limited situations. In this case, a game environment is a reasonable option for acquiring gunshots. For this reason, we constructed the BGG dataset, and this section covers the detailed description of our dataset.

A. Dataset Descriptions

We recorded the gun sounds by changing the type and position of guns to diversify distances and angles in the PUBG environment. As shown in Table III, the BGG dataset consists of 2,195 samples with 37 different types of guns and five directions, including a silence in which there is no gunfire, but noises exist. The distance from the firearms ranged from 0 meters to 600 meters. Audio was recorded in stereo (*i.e.,* two-channel audio), and each sample contains various environmental noises (*e.g.,* water splashing, walking, and bullet friction).

We constructed the two datasets from the collected sounds, namely BGG-CLS and BGG-LOC dataset. BGG-CLS was made for training and evaluating firearm classification models. Figure 1 shows the class distribution of BGG-CLS. The class denoted by *No gun* means that there is no gunfire and it only contains the sound of noise. We defined this class to prevent models from predicting the gun types in the absence of gunshots. Using BGG-LOC dataset, we can train and evaluate models that estimate the distance and direction of a gunshot. BGG-LOC dataset contains 1,024 sound samples categorized into three guns and *No gun*, and each sample is labeled with six distances (from 0 meters to 600 meters) and five directions (*front, back, left, right* and *center*). The distribution of each class is shown in Figure 2. Our dataset contains longer distances and more gun types than the existing datasets [5], [7]. To the best of our knowledge, there is no dataset for sound localization that addresses gunshots measured up to 600 meters.

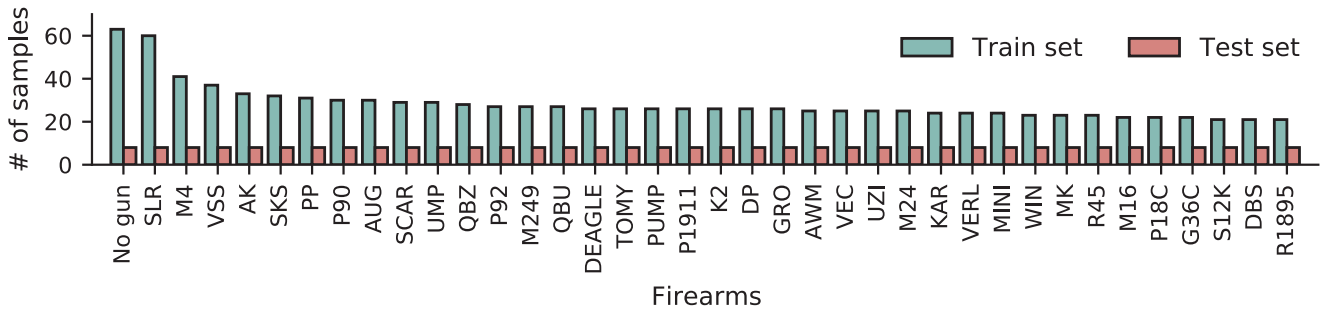

Fig. 1. The distribution of firearms in the BGG-CLS dataset is plotted. The blue and red bars indicate the number of train and test samples, respectively.

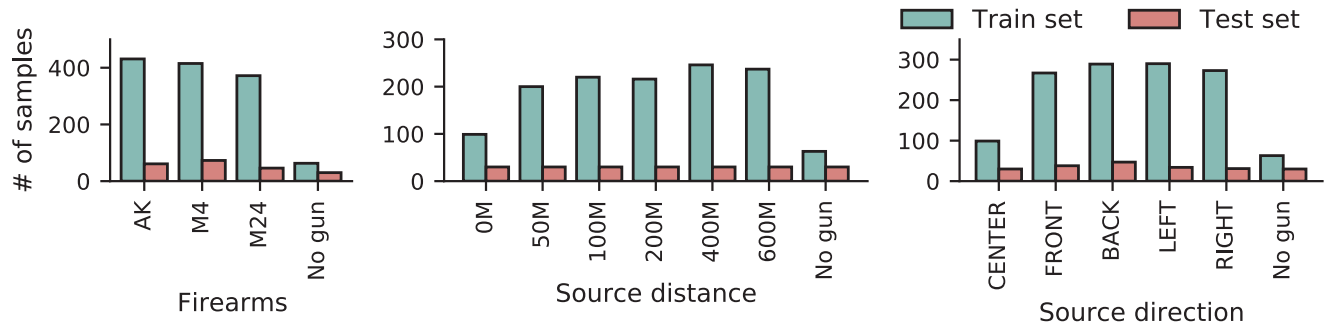

Fig. 2. The distributions of firearms, source distance, and source direction in the BGG-LOC dataset are plotted. The blue and red bars indicate the number of train and test samples, respectively.

B. Analysis of the Gun Sound

Figure 3 visualizes the audio samples in a two-channel waveform according to the gun type, distance, and direction. We checked whether the sound obtained from the game reflects the characteristics of the real-world sound. The volume of the gunshot decreases as the distance between an enemy and the player increases. At this time, the waveform of bullet friction with the air around the player (see the blunt shape in 600 meters in Figure 3) is more prominent than the waveform of gunshot (see the sharp shape in 0 meters). This phenomenon is observed when the player is far away from the enemy since the bullet friction sound is closer to the player than the location of the gunshot. In addition, the volume difference between the two channels appeared depending on the direction the player is facing and the location of the enemy's gunshot. Meanwhile, in the case of automatic weapon, the firing rate varies depending on the gun type. This difference can be seen in the spacing of the gunshots on the waveform. For example, firing rate of *M416* is 0.085 seconds, which is faster than *AKM* of which rate is 0.100 seconds. From the aforementioned observations, we assume that classifying firearms and localizing the enemy are possible from the gunshots obtained in the game environment.

IV. Method

This section formulates the sound classification and localization tasks, describes model architectures, and explains the training details.

A. Preliminaries

As a preliminary, we first define the notation of our dataset. Our BGG dataset is given as $\mathcal{D} = \{(x_n, y_n^g, y_n^s, y_n^d)\}_{n=1}^N$ where N indicates the number of instances. Here, $x_n \in \mathbb{R}^{T \times C}$ indicates a sound input where C denotes the number of channels and T represents the number of time steps. There exist three different types of labels; the one-hot representations of the sound type y_n^g, the discretized distance y_n^s, and the discretized direction y_n^d. Given a parameterized model f_θ, the goal of sound classification is to train the model which accurately predicts the sound type label y_n^g given x_n. Similarly, the goal of sound localization is to estimate the distance y_n^s and the direction y_n^d labels given x_n.

B. Model Architecture

Here, we describe the architectural details of our parameterized model f_θ. The model is divided into two components (i) feature extractor F_θ and (ii) classifier C_θ. For simplicity, we omit the instance subscript n throughout this section. First, an audio signal represented in a waveform $x \in \mathbb{R}^{C \times T}$ is converted to a spectrogram by Short-time Fourier transform (STFT) operation, which is widely used to detect how the frequency changes over time. Second, the processed spectrogram is passed on to the feature extractor F_θ to obtain a dense representation of the input sound. This module aggregates the spectrogram over the timestep.

$$z = F_\theta(\texttt{STFT}(x)) \tag{1}$$

where $z \in \mathbb{R}^D$ is aggregated along the time-axis, and D denotes the hidden dimension size.

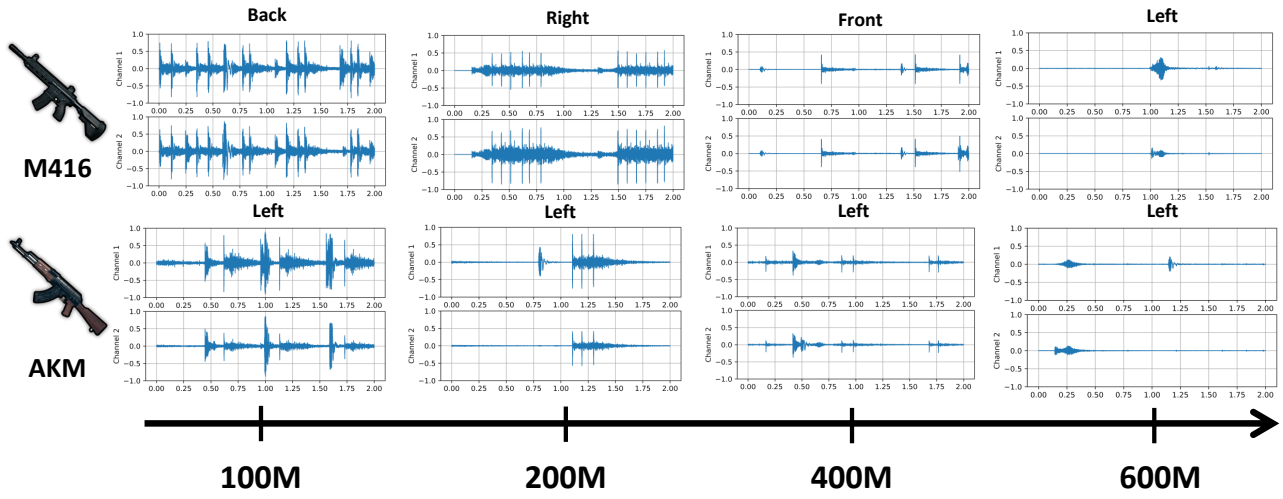

Fig. 3. Sample audio samples are visualized in the waveform. All samples are recorded in stereo. For each sound sample, the left and the right channel audio samples are plotted at the top and the bottom, respectively.

Finally, the extracted representation z is passed onto the classification module C_θ. The classification module consists of two fully-connected layers with a ReLU nonlinearity. Thereby, we obtain the predicted gunshot class probability p^g, distance class probability p^s, and and direction probability p^d.

$$p^g, p^s, p^d = C^g_\theta(z), C^s_\theta(z), C^d_\theta(z) \tag{2}$$

C. Training

For the gunshot classification dataset, `BGG-CLS`, the parameters θ are optimized to maximize the predicted gunshot class probability p^g of the ground truth gunshot label y^g. Cross-entropy loss is used to to maximize the log-likelihood.

$$\mathcal{L}_{cls} = -y^g \log p^g \tag{3}$$

For the gunshot localization dataset, `BGG-LOC`, the parameters θ are trained to maximize the predicted distance class p^s and direction class p^d of the ground truth distance label y^s and direction label y_d, respectively. Without loss of generality, we adopt cross-entropy loss to maximize the log-likelihood.

$$\mathcal{L}_{loc} = -(y^s \log p^s + y^d \log p^d) \tag{4}$$

V. Experiments

In this section, we describe the results of firearm classification and gunshot localization on `BGG` dataset. We evaluated three widely used models and two Transformer-based models [8], which are underexplored in firearm classification and gunshot localization tasks, to analyze the quality of our dataset.

A. Experimental Setting

We employed three sound classification methods: DCNNs [10], LSTM [22], and CRNNs [4]. DCNNs use stacked convolution layers where each layer is sequentially constructed with a 1D convolution operation, batch normalization, and nonlinearity functions. Convolutional layers can learn the temporal dependency of a spectrogram by sliding the filter of fixed size along with the time axis. LSTM is also widely used for learning temporal dependencies in sequential inputs while alleviating the vanishing gradient problem of RNNs via long-term and short-term memories. LSTM can be a suitable option for audio data whose time length increases according to a sampling rate compared to RNNs. We used bidirectional LSTM (bi-LSTM) in the experiment since the sound classification task has no restrictions on precedence in setting problems. CNNs are advantageous for learning local patterns, but have limitations in discovering long-distance temporal dependencies since they require more layers than RNNs. LSTMs are powerful in long-term dependency with fewer parameters than CNNs. Combining two methods, CRNNs showed considerable improvements in sound classification tasks. We additionally added a Transformer-based models. Transformer [8] has shown successful results by replacing CNNs and RNNs in the natural language processing, but has been relatively underexplored in firearm classification. The self-attention mechanism in Transformer is known to successfully learn dependencies between words in long sentences. Similar to CRNNs, we tested Transformer and CNN-Transformer.

We adopt two widely used classification metrics, accuracy (acc.) and F1 score, to measure the classification and localization performance. Accuracy is calculated by dividing the number of correctly predicted samples by the total number of samples. F1 score is measured by computing the harmonic mean of the precision and recall, and this measurement can also consider the class imbalance.

B. Experiments Results

We verified that deep learning models can predict firearms from in-game gunshots. Table II presents the performance of the five models trained on `BGG-CLS` dataset. The best accuracy and F1 score are reached by CNN-Transformer that uses convolutional layers as feature extractor and then

TABLE II
SOUND CLASSIFICATION RESULTS ON BGG−CLS DATASET THAT CONSISTS OF 37 GUN CLASSES EVALUATED WITH ACCURACY AND F1 SCORE.

Method	Validation		Test	
	Acc.	F1	Acc.	F1
Bi-LSTM	0.6627	0.5695	0.6068	0.5739
DCNN	0.8753	0.8381	0.8503	0.8315
CRNN	0.8445	0.8304	0.8279	0.8009
Transformer	0.6348	0.5571	0.5918	0.5642
CNN-Transformer	**0.9384**	**0.9196**	**0.9360**	**0.9290**

integrates the extracted features via Transformer. The DCNN and CRNN that show competitive performance with the best model also utilize convolutional layers as a feature extractor. On the other hand, Bi-LSTM and Transformer showed poor accuracy and F1 score. These results empirically indicate that local temporal patterns in the spectrogram are necessary to classify firearms.

For the gunshot localization task, we trained and evaluated the models on BGG−LOC to predict the distance and direction between the enemies and the game player. Since we formulated the localization task as the classification problem, the feature extractor of the models which learn discriminative features from the spectrogram is identical to the firearm classification models. The only difference is that there are two classifiers to predict the direction and distance. As shown in Table III, we confirmed that the overall performance of the DCNN, CRNN, and CNN-Transformer was higher than the Bi-LSTM and Transformer, showing the superiority of convolutional layers as a feature extractor. Note that the accuracy and F1 score of the models trained on BGG−LOC are higher than that of the models trained on BGG−CLS, because the number of firearm classes of BGG−CLS is smaller than that of BGG−LOC.

Unlike when predicting a firearm, Bi-LSTM and Transformer showed competitive performance with the DCNN, and the CRNN outperformed the DCNN when predicting distance. Figure 3 provides the possible explanation for this result. The enemies prefer aiming over continuous fire as the distance from the observer increases. Therefore, there are differences in the number of gunshots and the interval between them during the same time, depending on the distance. Although recurrent unit and self-attention mechanism can learn these temporal patterns, the max pooling layer of DCNN can ignore the differences because it only extracts large values along the time axis. Thus, we can reliably say that the temporal pattern is important as well as the difference in sound volume depending on the distance.

When predicting the direction of gunshots, DCNN and CNN-Transformer showed better performance than the other three methods. In the sight of the human auditory sense, the left and right ears correspond to two waveform channels and receive different audio signals depending on the direction of the sound due to the distance between the ears. In Figure 3, we observed the volume difference between the two channels. A max pooling after convolutional layers effectively learns these differences to predict direction. Finally, we confirmed that CNN-Transformer was the best method when comprehensively evaluating the three tasks, followed by the DCNN model. As a result, we designed a gunshot detection system in the game. The following section describes the qualitative results of the application.

C. Qualitative Result of the Application

This section demonstrates the qualitative results when our support system is deployed in the PUBG. We recorded the in-game video and extracted the audio signals from it. Given the extracted audio signal, our application estimates the type, distance and direction of gunshots according to time. The test scenario is as follows:

(a) For the first five seconds, a player is vigilant about his surroundings.
(b) An enemy at 100 meters to the right starts to fire an *AKM*.
(c) The player recognizes the attack and begins to strike back with an *M416* after facing the enemy.
(d) Finally, by judging that there is no chance of winning, the player runs away in the opposite direction of the enemy.

As shown in Figure 4, our application accurately detected the alternating *AKM* and *M416* gunshots in this scenario. Furthermore, in predicting the direction of gunshot, we confirmed that the application could detect the transition when the player turns his head to face the enemy (*i.e.,* from right to front). However, there is room for improvement in the distance estimation. According to the scenario, since the gunfires occur alternately at 100 meters and 0 meters, only the probability of 0 meters and 100 meters should be increased, but the probability of 50 meters increased.

D. Improving Performance in Real-World Benchmarks

A sufficient amount of data is necessary to achieve successful results using a deep learning model. However, the number of gunshot samples is smaller than the other sound samples in the existing datasets since gunshots occur less frequently than other sounds. Moreover, even if data is collected by firing guns in the real world, it is challenging to reproduce various situations and obtain detailed labeling due to the danger of firearms. Beyond the entertainment role, FPS games enable easily collecting data by simulating even dangerous situations in which a bullet passes by the side of the face. We show that the performance of gunshot-related tasks can be improved in real-world benchmarks through simple approaches using our BGG dataset. We reported the mean values measured in 10 experiments while mixing the training, validation, and test datasets.

The UrbanSound8K (Urban) dataset includes ten different classes of sounds, of which only one is the gunshot class. Compared to 8,358 samples belonging to classes other than gunshots, the number of samples included in the gunshot class is the smallest at 374, resulting in the lowest classification accuracy. We compared models trained on the urban dataset additionally containing gunshots from the BGG as the samples of the urban gunshot class with models trained by simply

TABLE III

GUNSHOT LOCALIZATION RESULTS OVER BGG−LOC DATASET THAT CONSISTS OF THREE GUN CLASSES, SIX DIRECTIONS, AND SEVEN DISTANCES.

Method	Rank	Firearm		Distance		Direction	
		Acc	F1	Acc	F1	Acc	F1
Bi-LSTM	5	0.8717 ± 0.0210	0.8776 ± 0.0190	0.8593 ± 0.0189	0.8596 ± 0.0169	0.7143 ± 0.0274	0.7257 ± 0.0274
DCNN	2	0.9507 ± 0.0107	**0.9424** ± 0.0115	0.8668 ± 0.0129	0.8695 ± 0.0121	0.8792 ± 0.0065	**0.9099** ± 0.0112
CRNN	3	0.9339 ± 0.0112	0.9292 ± 0.0112	0.8901 ± 0.0150	0.8912 ± 0.0149	0.7900 ± 0.0738	0.7789 ± 0.0681
Transformer	4	0.9011 ± 0.0131	0.9015 ± 0.0124	0.8327 ± 0.0162	0.8307 ± 0.0159	0.7815 ± 0.0214	0.8079 ± 0.0184
CNN-Transformer	1	**0.9529** ± 0.0054	0.9386 ± 0.0075	**0.9150** ± 0.0165	**0.9156** ± 0.0159	**0.9313** ± 0.0085	0.9003 ± 0.0132

TABLE IV

THE RESULT OF SOUND CLASSIFICATION OVER URBAN DATASET EVALUATED WITH ACCURACY AND F1 SCORE. WE REPORT THE MEAN OF THE VALUES OBTAINED FROM 10 CASES.

Method		Urban only → Urban + BGG Valid	Urban only → Urban + BGG Test
DCNN	Acc.	0.7556 → **0.7748**	0.7370 → **0.7736**
	F1	0.7542 → **0.7718**	0.7206 → **0.7650**
	Gun acc.	0.5747 → **0.6272**	0.2661 → **0.4708**
CRNN	Acc.	0.7456 → **0.7616**	0.7386 → **0.7680**
	F1	0.7442 → **0.7615**	0.7158 → **0.7567**
	Gun acc.	0.4600 → **0.5051**	0.1998 → **0.3841**
CNN-Transformer	Acc.	0.7726 → **0.7963**	0.7693 → **0.8006**
	F1	0.7696 → **0.7946**	0.7495 → **0.7951**
	Gun acc.	0.5661 → **0.6534**	0.2709 → **0.5249**

augmenting the gunshot samples in the existing urban by applying random cropping and speed-changing methods. For fair comparison, the number of added data was set to the same number for both models. Table IV shows that the accuracy and F1 score are improved with a significant gap when the BGG is used. The large gap in Gun acc. means that the performance improvement was attributed to the model's ability to distinguish gun sounds by adding the BGG.

To validate in-game data in gunshot classification and localization tasks, we trained and evaluated the baseline models on the Forensic (Foren) dataset, consisting of 2,241 samples categorized into 18 different types of firearms, eight distances, and 13 directions. Unlike the Urban dataset, which also contains non-gunshots, the Foren dataset contains only gunshots. After training models that predict firearms, distance, and direction on the BGG dataset, we use the parameters of the feature extractor as the initial parameters of the same network classifying firearms, distance, and direction into predefined classes of the Foren. We compared the models that were pretrained on the BGG and those that were only trained on the Foren from scratch. As shown in Table V, the models trained on the BGG outperformed the models trained from scratch in accuracy and F1 score. Although the BGG dataset and the Foren dataset contain different firearms, the parameters pretrained with the in-game gunshots in the BGG dataset strengthen the ability of the model to classify the real-world gunshots into firearms from the Foren dataset. In the case of predicting direction and distance, improvement in accuracy when the BGG is used is not significant because the range of distance and direction

TABLE V

THE RESULTS OF GUNSHOT CLASSIFICATION AND LOCALIZATION OVER FOREN DATASET EVALUATED WITH ACCURACY AND F1 SCORE. WE REPORT THE MEAN OF THE VALUES OBTAINED FROM 10 CASES.

Firearm		Foren only → Foren + BGG Valid	Foren only → Foren + BGG Test
DCNN	Acc.	0.6696 → **0.6968**	0.6832 → **0.7016**
	F1	0.6154 → **0.6518**	0.6200 → **0.6509**
CRNN	Acc.	0.6181 → **0.6692**	0.6096 → **0.6616**
	F1	0.5856 → **0.6386**	0.5726 → **0.6236**
CNN-Transformer	Acc.	0.6741 → **0.7260**	0.6798 → **0.7410**
	F1	0.6362 → **0.7082**	0.6346 → **0.7126**

Distance		Foren only → Foren + BGG Valid	Foren only → Foren + BGG Test
DCNN	Acc.	0.6351 → **0.6611**	0.6691 → **0.6941**
	F1	0.5053 → **0.5319**	0.5148 → **0.5500**
CRNN	Acc.	0.7501 → **0.7836**	0.7763 → **0.7946**
	F1	0.6732 → **0.6995**	0.6988 → **0.7049**
CNN-Transformer	Acc.	0.8429 → **0.8526**	0.8566 → **0.8641**
	F1	0.8192 → **0.8538**	0.8323 → **0.8558**

Direction		Foren only → Foren + BGG Valid	Foren only → Foren + BGG Test
DCNN	Acc.	0.7163 → **0.7560**	0.7820 → **0.8016**
	F1	0.5970 → **0.6319**	0.6266 → **0.6722**
CRNN	Acc.	0.7402 → **0.7439**	0.8426 → 0.8033
	F1	0.6898 → **0.6999**	0.7280 → 0.7188
CNN-Transformer	Acc.	0.8628 → **0.8713**	0.8607 → **0.8827**
	F1	0.8141 → **0.8376**	0.8421 → 0.8420

covered by the Foren and the BGG is different. Nevertheless, except for three cases, all performance has increased. Thus, we can say that it can be used to create a support system in the real world by using in-game data.

VI. CONCLUSION

Recognizing enemies from a gunshot has substantial benefits to prepare for the threat in FPS games and the real world. In the real world, the support system that identifies and locates gunshots plays a vital role in protecting social safety from the misuse of guns. However, due to the rarity and unpredictable occurrence of gunshots in the real world, the number of gun sound samples is insufficient to build deep learning-based gunshot classification and localization models.

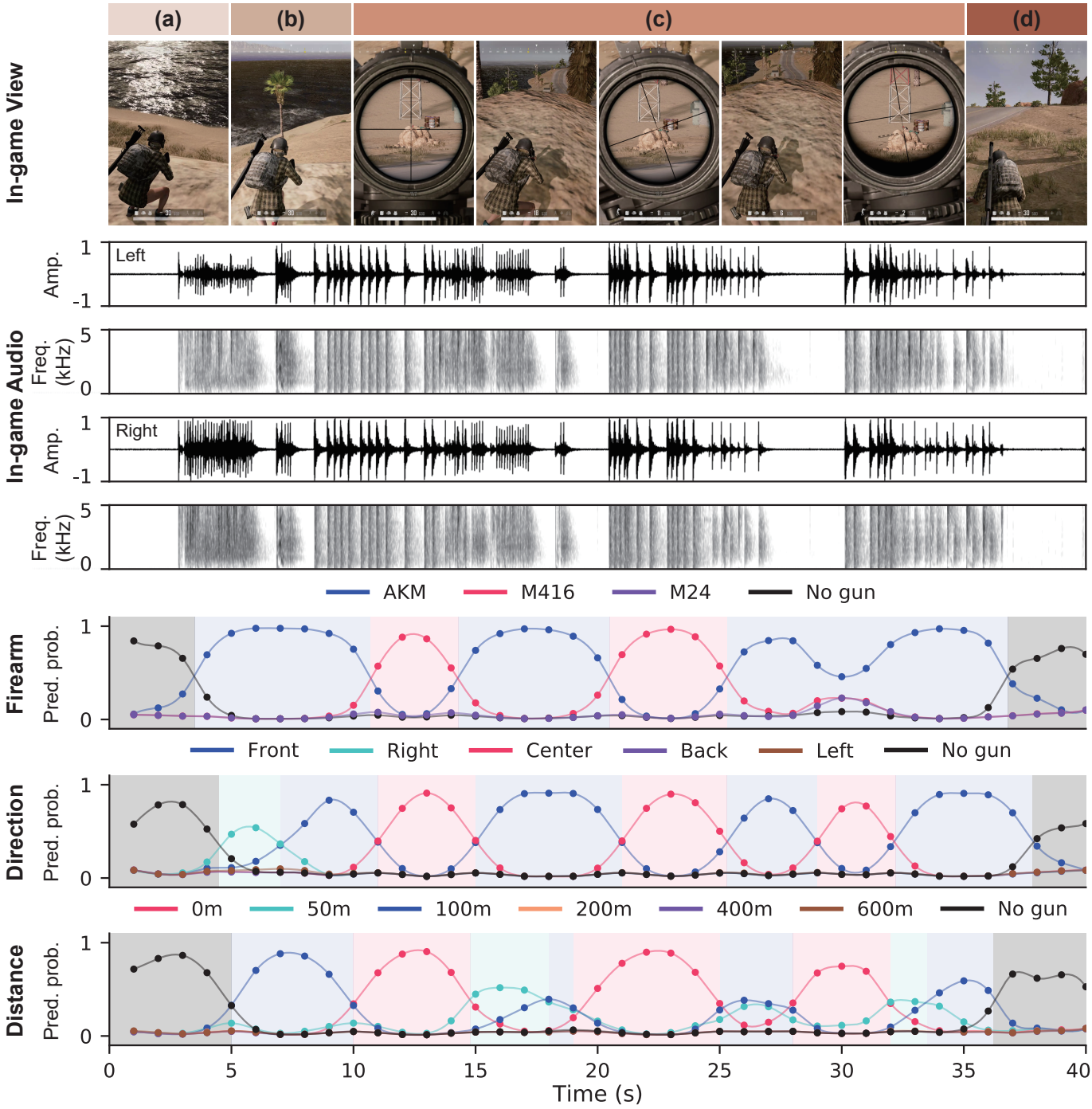

Fig. 4. Demonstration of our gunshot classification and localization application on the PUBG environment. (**a**), (**b**), (**c**), and (**d**) refer to each phase in the confrontation with the enemy. Along with in-game screenshots, a sound sample recorded from the game is plotted in the raw waveform and the spectrogram. Three rows at the bottom visualize the prediction results from firearm classification and gunshot localization model. The visualizations are highlighted based on the prediction probabilities. At the beginning and the end of the engagement, no gunshot occurs in the environment (**a** and **d**). The player turns right to find the enemy when the enemy starts to attack the player (**b**), and the predicted direction changes from right to front (highlights changes from *sky-blue* to *dark-blue*). During engagement (**c**), the player and the enemy fire their firearms alternatively; therefore, we can see all predictions aligned (highlights changes between *dark-blue* and *red*).

As an alternative, by recording gunshots in an FPS game, we constructed a gunshot dataset for the firearm classification and localization tasks. We validated the practicality of the in-game gunshot data by analyzing the deep-learning models trained on our BGG dataset. Moreover, we demonstrated that the gun sound dataset collected from the game improves real-world gunshot classification and localization accuracy.

ACKNOWLEDGEMENTS

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No.2019-0-00075, Artificial Intelligence Graduate School Program (KAIST)) and the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. NRF-2022R1A2B5B02001913).

REFERENCES

[1] M. Yiwere and E. J. Rhee, "Distance estimation and localization of sound sources in reverberant conditions using deep neural networks," *Int. J. Appl. Eng. Res*, vol. 12, no. 22, pp. 12 384–12 389, 2017.

[2] ——, "Sound source distance estimation using deep learning: an image classification approach," *Sensors*, vol. 20, no. 1, p. 172, 2020.

[3] K. Choi, G. Fazekas, M. Sandler, and K. Cho, "Convolutional recurrent neural networks for music classification," in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2017, pp. 2392–2396.

[4] E. Cakır, G. Parascandolo, T. Heittola, H. Huttunen, and T. Virtanen, "Convolutional recurrent neural networks for polyphonic sound event detection," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 25, no. 6, pp. 1291–1303, 2017.

[5] J. F. Gemmeke, D. P. W. 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)*, New Orleans, LA, 2017.

[6] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[7] S. Raponi, I. Ali, and G. Oligeri, "Sound of guns: digital forensics of gun audio samples meets artificial intelligence," *arXiv preprint arXiv:2004.07948*, 2020.

[8] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[9] K. Nakamura, K. Nakadai, F. Asano, Y. Hasegawa, and H. Tsujino, "Intelligent sound source localization for dynamic environments," in *2009 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2009, pp. 664–669.

[10] N. Yalta, K. Nakadai, and T. Ogata, "Sound source localization using deep learning models," *Journal of Robotics and Mechatronics*, vol. 29, no. 1, pp. 37–48, 2017.

[11] H.-Y. Lee, J.-W. Cho, M. Kim, and H.-M. Park, "Dnn-based feature enhancement using doa-constrained ica for robust speech recognition," *IEEE Signal Processing Letters*, vol. 23, no. 8, pp. 1091–1095, 2016.

[12] J. Stachurski, L. Netsch, and R. Cole, "Sound source localization for video surveillance camera," in *2013 10th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2013, pp. 93–98.

[13] P.-A. Grumiaux, S. Kitić, L. Girin, and A. Guérin, "A survey of sound source localization with deep learning methods," *arXiv preprint arXiv:2109.03465*, 2021.

[14] R. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, no. 3, pp. 276–280, 1986.

[15] C. Knapp and G. Carter, "The generalized correlation method for estimation of time delay," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 24, no. 4, pp. 320–327, 1976.

[16] j. h. hansen and h. boril, "gunshot detection systems: methods, challenges, and can they be trusted?" *journal of the audio engineering society*, october 2021.

[17] T. Ahmed, M. Uppal, and A. Muhammad, "Improving efficiency and reliability of gunshot detection systems," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013, pp. 513–517.

[18] J. Bajzik, J. Prinosil, and D. Koniar, "Gunshot detection using convolutional neural networks," in *2020 24th International Conference Electronics*, 2020, pp. 1–5.

[19] R. Baliram Singh, H. Zhuang, and J. K. Pawani, "Data collection, modeling, and classification for gunshot and gunshot-like audio events: a case study," *Sensors*, vol. 21, no. 21, p. 7320, 2021.

[20] H. Lim, J. Park, and Y. Han, "Rare sound event detection using 1d convolutional recurrent neural networks," in *Proceedings of the detection and classification of acoustic scenes and events 2017 workshop*, 2017, pp. 80–84.

[21] K. Shimada, Y. Koyama, and A. Inoue, "Metric learning with background noise class for few-shot detection of rare sound events," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 616–620.

[22] I. Lezhenin, N. Bogach, and E. Pyshkin, "Urban sound classification using long short-term memory neural network," in *2019 federated conference on computer science and information systems (FedCSIS)*. IEEE, 2019, pp. 57–60.