Article

Source Type Classification and Localization of Inter-Floor Noise with a Single Sensor and Knowledge Transfer between Reinforced Concrete Buildings [†]

Hwiyong Choi ¹ , Woojae Seong ^{1,2} and Haesang Yang ^{1,*}

¹ Department of Naval Architecture and Ocean Engineering, Seoul National University, Seoul 08826, Korea; its_me_chy@snu.ac.kr (H.C.); wseong@snu.ac.kr (W.S.)

² Research Institute of Marine Systems Engineering, Seoul National University, Seoul 08826, Korea

* Correspondence: coupon3@snu.ac.kr; Tel.: +82-2-880-8493

[†] This paper is an extended version of our paper published in 28th European Signal Processing Conference, Amsterdam, The Netherlands on 24 August 2020.

Abstract: A convolutional neural network (CNN)-based inter-floor noise source type classifier and locator with input from a single microphone was proposed in [*Appl. Sci.* **9**, 3735 (2019)] and validated in a campus building experiment. In this study, the following extensions are presented: (1) data collections of nearly 4700 inter-floor noise events that contain the same noise types as those in the previous work at source positions on the floors above/below in two actual apartment buildings with spatial diversity, (2) the CNN-based method for source type classification and localization of inter-floor noise samples in apartment buildings, (3) the limitations of the method as verified through several tasks considering actual application scenarios, and (4) source type and localization knowledge transfer between the two apartment buildings. These results reveal the generalizability of the CNN-based method to inter-floor noise classification and the feasibility of classification knowledge transfer between residential buildings. The use of a short and early part of event signal is shown as an important factor for localization knowledge transfer.

Keywords: Inter-floor noise; single sensor acoustics; convolutional neural network; knowledge transfer

Citation: Choi, H.; Seong, W.; Yang, H. Source Type Classification and Localization of Inter-Floor Noise with a Single Sensor and Knowledge Transfer between Reinforced Concrete Buildings. *Appl. Sci.* **2021**, *11*, 5399. <https://doi.org/10.3390/app11125399>

Academic Editor: Edoardo Piana

Received: 21 May 2021

Accepted: 7 June 2021

Published: 10 June 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

1.1. Motivation

In multi-dwelling units, noises generated by occupants propagate through the structures and exert an unpleasant effect on neighboring occupants [1–4], which is a serious problem in major cities in Korea, where most residential buildings are multi-dwelling units [5]. For example, 62% of the residential buildings of South Korea are classified as apartment buildings of more than five storeys [6]. Accordingly, the Floor Noise Neighborhood Center of the Korea Environment Corporation [7], affiliated with the Ministry of Environment, received 152,061 complaints about inter-floor noise from 2012 to 2019 [8].

It is challenging to identify the inter-floor noise traveling through multi-storey residential buildings owing to the human ears' failures to intercept these sounds. Incorrect identification of inter-floor noise by human hearing often causes conflicts among occupants. Some conflicts originated from wrong identification of inter-floor noise. In one case, the victim complains to an occupant who did not generate any inter-floor noise. In other case, the person who made the noise pretends not to know about it and ignores the victim's complaints. For both cases, technical identification of inter-floor noise can provide a proper basis for settling the dispute. Human ears process and discriminate sound accurately, but identification of inter-floor noise via machine could provide less biased results. In the authors' previous studies [9,10], a single microphone-based inter-floor noise source type

classifier/locator was proposed to assist the identification problem using convolutional neural network (CNN)-based supervised learning, rather than approaches with multiple channels of accelerometers [11,12] or geophones [13] networks. The method was verified on an actual dataset obtained from a campus building. This method can be implemented in a personal mobile phone device and constructs a data-driven model that helps reduce failure of noise identification by the human ears with less human bias and provide a proper basis for settlement in the case the offender disregards the complaints. However, validation of the generalizability of the method in actual residential buildings was left for future study [10]. Therefore, the method needs to be verified for many scenarios and to determine its limitations.

1.2. Related Literature

Sound classification, which deals with tasks that are similar to the source type classification in this study, has been studied in acoustic scene classification (ASC) fields [14,15]. The conventional data-driven methods for ASC extract features from audio waveforms adopting Mel-filterbanks, Mel-frequency cepstral coefficient (MFCC), or principal component analysis (PCA) and classify the extracted features into a category via majority voting or support vector machine (SVM) [16–18]. Recently proposed methods follow the deep neural network (DNN)-based scheme after adoption of CNN [19].

In the conventional acoustic source localization, a method based on the triangulation technique (Tobias algorithm) obtained analytical solutions using multiple channels of sensors on known environmental properties, such as positions of sensors and sound speed [20]. The triangulation technique typically used the time of arrival (TOA) and the group velocity of the direct wave [21]. A pair of TOAs, the absolute time instants when a transmitted signal is detected by each sensor, provides three equations of the circle in two dimensions. Solution for source position exists in the form of the intersection of the three circles. Complementary triangulation techniques were introduced to minimize the effect of the dispersive guided waves. They were realized by optimizing the error function [22] or combining the continuous wavelet transform, Newton's method, and the line search algorithm [23]. A model-based impact locator with a single sensor was proposed and validated in a plate structure [24].

Learning-based methods are possible alternatives to the model-based methods. These approaches learn the relationship between the given source positions and signals traveled through a structure. Grabec et al. [25] adopted the adapted learning and series of reference signals with position information to localize sources. Kosel et al. [26] adopted the responses of discrete sources to train the locator. As an application to the human-device interface, Ing et al. [27] introduced the time-reversal process for localizing the impact of a finger on a plate. Ciampa et al. [21] demonstrated the feasibility of adopting the time-reversal process for localization in a composite structure. Ruiz et al. [28] proposed an impact localization method based on projections to latent structures. Although these methods require data gathering, they do not need knowledge on sound speed of the medium or receiver position.

Neural network (NN) is an important tool for approximating the complex relationship between response signals and their source types/positions. In [29], the localization of impact and damage detection was demonstrated simultaneously using a multilayer perceptron. In [30], a CNN-based approach was introduced to localize acoustic sources in a plate with rivet-connected stiffeners. Notably, a few previous studies [21,27,30] adopted a single sensor and demonstrated the feasibility of single sensor-based localization. However, these approaches were analyzed in plate-like or simplified composite plates, which are considered to be simpler structures than those in real life.

For practical application, it is important to elucidate the feasibility of the algorithm using actual datasets. Previous studies on single sensor-based localization in real-world applications have been conducted. A previous study [31] demonstrated localization in a room via echo labeling. Two other studies [32,33] synthesized training data via model-

based simulations to train their CNNs. The range between a source and single hydrophone on actual sea trial data were presented. In addition, the former [32] verified the practicality of depth estimation, whereas the later [33] focused on classifying ocean bottom type.

Localization in actual buildings has been studied in indoor occupant localization fields. Considering the dispersive nature of waves in a plate and sign of the measured time differences of arrival (SO-TDOA), Bahroun et al. [11] introduced a foot step localization technique in a damped and dispersive media. Poston et al. [12] proposed a footstep type aware footstep localization technique, which identifies a given footstep type (compression or non-compression) and applies type-wise localization algorithm. In addition, a tracking algorithm for moving occupants on linear trajectories was studied [34]. Mirshekari et al. [13] demonstrated the limitation of signal distortion to enhance the TDOA estimation using wavelet transform for the localization of footstep-induced vibrations. Woolard [35] studied a learning-based event localization in a hallway and on stairs in a campus building using a nearest neighbor algorithm with three accelerometers. In addition, the feasibility of a single sensor-based direction of arrival estimation on a beam was demonstrated via simulation.

1.3. Approach

This study presents the CNN-based source type classifier and locator with a single microphone on inter-floor noise data obtained from two actual apartment buildings to verify the generalizability of the method, which considered important for data-driven approach. In addition, the feasibility to learned source type and localization knowledge transfer between the similar reinforced concrete buildings is presented. This approach significantly relies on deep learning to formulate the source type classifier and locator for air-concrete-steel mixed environments, where building properties are insufficiently known, as well as for structure with high structural complexity in the acoustic medium. Similar to the learning-based method proposed in the previous studies [9,10], it learns responses with source type/position labels transmitted from discrete positions in the buildings to formulate the data-driven identification of inter-floor noise in reinforced concrete building using a single microphone, thereby extending the application of deep learning. Accordingly, inter-floor noise was obtained from two actual apartment buildings. The data points selected were on slabs of allowed rooms for experiments, whereas the campus building dataset (SNU-B36-50E [10]) in the previous study includes those on corridor slabs in the two-dimensional spaces alone. Several inter-floor noise identification and knowledge transfer tasks were conducted on the new dataset to demonstrate the generalizability, and to elucidate the limitations and uncertainties of the method.

1.4. Contributions

The contributions of this study are summarized as follows. (1) Inter-floor noise datasets were built with noise samples obtained from two actual reinforced concrete apartment buildings to study data-driven source type classification and localization in actual reinforced concrete buildings. (2) The CNN-based source type classification and localization with a single microphone was demonstrated via several tasks on the datasets. In addition, the limitations of this approach were discussed. (3) The feasibility of the learned source type and localization knowledge transfer between the apartment buildings was demonstrated. Provided the source type and localization knowledge of trained samples can be reused for tasks, without required training or even under data sparsity, the noise identification method can be used widely. (4) It was empirically shown that using a short and early parts of an inter-floor noise signal is effective for localization knowledge transfer between the buildings.

The remainder of the paper is organized as follows. The apartment building inter-floor noise datasets are explained in Section 2. An onset detection is described in Section 3 that finds the event start position of an inter-floor noise signal to reduce human effort to achieve visual annotation of the event. Several tasks for verifying the source type classifier, locator, and knowledge transfer between two apartment buildings are prepared. The measured

performance of the approach is reported and discussed in Section 4. Finally, the paper is summarized in Section 5.

2. Apartment Building Inter-Floor Noise Datasets

The two datasets adopted in this study contain inter-floor noise recorded using a single microphone in two actual apartment buildings. They are designed to study the CNN-based source type classification and localization in apartment buildings. These extend the dataset obtained from the campus building in the previous study [10]. The data points were selected to simulate inter-floor noises on the floor above/below based on the noise statistics [4], which provides the main unpleasant noise types and source positions to occupants. The key purposes of the dataset are for verifying (1) the generalizability of CNN for source type classification and localization on inter-floor noise in actual reinforced concrete buildings not only in the campus building as exhibited in the previous work [10]; (2) source type classification and localization of inter-floor noise transmitted through unlearned floor sections and from unlearned positions, which can be seen as knowledge transfer within a single building; and (3) source type and localization knowledge transfer between two similar reinforced concrete buildings.

Selecting source type and position of inter-floor noise for data construction was discussed sufficiently in the previous study [10] based on the noise statistics [4]. The statistics provides the source types and positions of the identified inter-floor noises from the analysis of the 119,500 complaints investigated by the center from 2012 to March of 2018. The identified source types and their contributions to inter-floor noise complaints are attributed to footsteps (71.0%), hammering (3.90%), furniture (3.3%), home appliances (vacuum cleaner, laundry machines, and television) (3.3%), doors (2.0%), and unidentified or unrecorded sources (10.1%). Of the identified inter-floor noises, 79.4% were from the floor above and 16.3% were from the floor below. In other words, 95.7% of the complaints originated from inter-floor noises on the floors above/below.

Inter-floor noises were generated in two apartment buildings referring to the discussion. The inter-floor noise obtained from the two apartment buildings can be classified into five source types, as shown in Figure 1. They are the same source types as those included in the dataset from the campus building: a medicine ball falling to the floor from a height of 1.2 m (MB), a hammer dropped from a height of 1.2 m above the floor (HD), hammering (HH), dragging a chair (CD), and operating a vacuum cleaner (VC). The inter-floor noise generating procedures are the same as those in the previous work [10].

Figure 1. Five noise types included in the datasets. Reprinted [10].

Figures 2 and 3 present floor plans and building elevations of apartment buildings I (APT I) and II (APT II), respectively, where the inter-floor noise was obtained. The circles (○) and squares (□) represent the positions of the noise sources and receivers, respectively, and the labels near the circles indicate the source positions. These apartment buildings are reinforced concrete-frame structures and partitioned with concrete walls. Such building structures are reported as the most widely used types for modern buildings in South Korea [36,37]. The slabs of the apartment buildings were covered with vinyl flooring. The construction details of APT I and APT II are a reinforced concrete wall and reinforced concrete masonry structure [38], respectively. The reinforced concrete wall structure withstands the load generated by its own weight by wall and has been the mainstream of modern apartment building construction in South Korea. There is a statistics

provided by the Korean Ministry of Land, Infrastructure and Transport that 98.5% of new residential multi-dwelling units during 2007–2017 were constructed using this method [39]. The reinforced concrete masonry structure was usually adopted for the construction of low-rise buildings during the 1960s–1980s in South Korea.

Both APT I and APT II datasets are designed to obtain the five source types of inter-floor noise from the floors above/below, similar to the campus building dataset. Inter-floor noise was recorded using a smartphone [40] microphone with a sampling rate f_s of 44,100 Hz. The duration of each recording is approximately 5 s, and each recording contains a single event. The height of the receiver was 1.5 m above the floor, as set for the campus building dataset. Both datasets could be split into a training/validation and test dataset.

Figure 2. Noise source (○) and receiver (□) positions used for inter-floor noise generation at APT I. (a) Floor plan. (b) Building elevation 1 (APT 1). (c) Building elevation 1' (APT 1').

Figure 3. Noise source (○) and receiver (□) positions used for inter-floor noise generation at APT II. (a) Floor plan. (b) Building elevation 2 (APT 2). (c) Building elevation 2' (APT 2').

Obtaining a training/validation dataset from APT I was prepared as follows. The five source types were generated at 1-A and 1-B, and sampled with the receiver on the floor above/below as illustrated in Figure 2b. The data points 1-A and 1-B are the centers of the two spaces evenly dividing the room allowed for the experiment. Their positions relative to the receiver are labeled as

$$\mathcal{Y}_{p,APT1} = \{1-A-a, 1-B-a, 1-A-b, 1-B-b\}, \quad (1)$$

where a and b represent the floors above and below relative to the receivers, respectively. VC from the floor below, i.e., VC from 3 F to 4 F, was not recorded, as this source type was barely audible from the floor above (4 F). For obtaining a test dataset from APT I, inter-floor noise was generated at 1-A', 1-B', 1-C, 1-D, and 1-E, as illustrated in Figure 2c. The noise was sampled with the receiver on the floor below (3 F). 1-A' and 1-B' are at the same XY positions as those of 1-A and 1-B, respectively. Therefore, 1-A'-a and 1-B'-a can be considered the same as 1-A-a and 1-B-a from the viewpoint of the receivers. The five source types were generated at these two positions (1-A' and 1-B'). In addition, MB and HH were generated at 1-C, 1-D, and 1-E, where these source types occupy large portion of the identified source types in the complaint analysis [4]. Their positions relative to the receiver position are labeled as

$$\mathcal{Y}_{p,APT1'} = \{1-A'-a, 1-B'-a, 1-C-a, 1-D-a, 1-E-a\}. \quad (2)$$

Union of these two separately labeled data domains, $\mathcal{D}_{\text{APT I}} = \{\mathcal{X}, \mathcal{Y}_{\text{APT I}}\}$ and $\mathcal{D}_{\text{APT I}'} = \{\mathcal{X}, \mathcal{Y}_{\text{APT I}'}\}$, are combined and represented as $\mathcal{D}_{\text{APT I}} = \{\mathcal{X}, \mathcal{Y}_{\text{APT I}}\}$. Most multi-storey residential buildings have almost the same structure for all floors. However, the deployment of goods, such as furniture can differ on each of the floor and act as uncertainties. The test dataset in $\mathcal{D}_{\text{APT I}'} = \{\mathcal{X}, \mathcal{Y}_{\text{APT I}'}\}$ can be adopted to verify the robustness of a source type classifier or locator against these scenarios.

The source types, except VC, were generated at all noise source positions in APT II. Obtaining a training/validation dataset (APT 2 dataset) from APT II was prepared as follows. The four source types were generated at 2-A, 2-B, 2-C, and 2-D and sampled with the receivers on the floor above/below as illustrated in Figure 3b. Their positions relative to the receiver position are labeled as

$$\mathcal{Y}_{\text{p, APT 2}} = \{2\text{-A-a}, 2\text{-B-a}, 2\text{-C-a}, 2\text{-D-a}, 2\text{-A-b}, 2\text{-B-b}, 2\text{-C-b}, 2\text{-D-b}\}. \quad (3)$$

Each data point is at the center of each room or space, e.g., 2-A and 2-C are at the center of the living room and bed room. For obtaining a test dataset (APT 2' dataset) from APT II, inter-floor noise was generated at 2-A', 2-B', 2-C', and 2-D', as illustrated in Figure 3c. They are at the same XY positions as those for the training/validation dataset. Their positions relative to the receiver position are labeled as

$$\mathcal{Y}_{\text{p, APT 2}'} = \{2\text{-A'-a}, 2\text{-B'-a}, 2\text{-C'-a}, 2\text{-D'-a}\}. \quad (4)$$

Union of $\mathcal{D}_{\text{APT 2}} = \{\mathcal{X}, \mathcal{Y}_{\text{APT 2}}\}$ and $\mathcal{D}_{\text{APT 2}'} = \{\mathcal{X}, \mathcal{Y}_{\text{APT 2}'}\}$ is represented as $\mathcal{D}_{\text{APT II}} = \{\mathcal{X}, \mathcal{Y}_{\text{APT II}}\}$.

APT I dataset contains noise samples generated only on the living room slabs. Approximately 50 inter-floor noise events were obtained for source type at each relative position in APT I. Inter-floor noise was generated on the bedroom slabs as well as on the living room slabs in APT II. Approximately, 60 inter-floor noise events were obtained for each source type at each relative position. APT I and APT II datasets were obtained in each building per day. The total numbers of the inter-floor noise events in APT I and APT II dataset are 1785 and 2880, respectively. The data points in APT II may have more generalized conditions than those in APT I, because they are distributed in the wider three-dimensional spaces.

3. Inter-Floor Noise Classification

3.1. Onset Detection

In the previous studies [9,10], a visually annotated inter-floor noise sample with a duration of 3 s was converted to a log-scaled Mel-spectrogram and classified into a source type and position category by a CNN-based classifier. However, the human visual annotation requires effort to annotate large amounts of data and knowledge.

In the seismic signal-processing field, automatic seismic event-detection algorithms have been developed to replace laborious visual detection by humans. Allen [41] developed an earthquake timing-detection algorithm using time averaging and zero-crossing rate measurement of signals over seismometers. Allen's algorithm was set as a baseline in many other onset-picking studies. A modified onset-picking method was developed in the acoustic emission field using the Akaike Information Criterion (AIC) [42], referring to Allen's algorithm and the extended work [43]. Applications of this approach in the acoustic emission field can be found [12,44]. A broad-band maximum-likelihood method to estimate parameters and detect seismic events was studied [45]. Higher-order statistics (HOS)-based onset-picking methods were compared with Allen's algorithm and analyst's picks [46,47]. These HOS-based methods were simple in implementation, showed similar results to analysts' picks, and demonstrated less detection failure under high noise levels than that of Allen's algorithm.

Most of the signals in the inter-floor noise datasets are impact signals whose duration times are short and show a drastic energy rise at the event start positions. In this study, an onset-detection method using kurtosis is employed to detect the onset of the inter-floor noise signals. The kurtosis measures the heaviness of the tails representing the non-Gaussianity. This method was selected for the following reasons: (1) using a single HOS-property, kurtosis, is compact in implementation; (2) there exist very small differences between the onset detection results using skewness and kurtosis [46]; (3) the signals adopted in this study are obtained with a single sensor, so that exact onset-time picking is not required, unlike methods employing multiple channels of sensors to improve the performance of calculated properties, e.g., TDOA; and (4) CNN comprises shift-invariance in time axis in its input [48]. Equation (5) estimates the kurtosis of the M -sample sliding window, which returns a maximum value of $\hat{\gamma}_4(k)$, ($k = M, M + 1, \dots, L$) at an onset position for a given L -sample signal vector $\mathbf{s}(m)$ [47].

$$\hat{\gamma}_4(k) = (M - 1) \frac{\sum_{m=k-M+1}^k (\mathbf{s}(m) - \hat{a}_k)^4}{\left[\sum_{m=k-M+1}^k (\mathbf{s}(m) - \hat{a}_k)^2 \right]^2} - 3, \quad (5)$$

$$\text{where } \hat{a}_k = \frac{1}{M} \sum_{m=k-M+1}^k \mathbf{s}(m).$$

A confidence interval is usually considered to bound and ensure the probability of the signal's existence [47], which may be required in actual applications to reduce false alarms. However, this was not considered in this study, because the audio clips in the datasets do contain a single event such that a position with the maximum $\hat{\gamma}_4(k)$ can be assumed as an onset. The window size M is usually selected to achieve the minimum difference between the analyst's pick and an estimated onset position. $M = 3000$ is sufficient to find an onset position.

3.2. Convolutional Neural Network-Based Classifier

The CNN-based classifiers designed for image recognition already demonstrated source type classification and localization in the previous studies [9,10]. These prior studies extended the application of CNN to source type classification and localization of inter-floor noise. Several state-of-the-art CNNs [49–51], which have been widely employed in many other applications, were tested against inter-floor noise classification tasks using a dataset obtained from the campus building. The inter-floor noise classification tasks to be stated and solved in this study proceeded upon the assumption that CNNs designed for image recognition can be adopted for these tasks.

In this study, VGG16 [51] is employed as a CNN-based feature extractor from an inter-floor noise in the form of log-scaled Mel-spectrogram $\mathbf{P}$. This architecture showed the best performance among several state-of-the-art CNNs for inter-floor noise classification in the previous study [10]. An alternative to this approach is a one-dimensional CNN, which does not require conversion of the inter-floor noises to image-like features. The kernels of one-dimensional CNN eventually learn a set of Mel-filters, which can be considered as a set of basis functions in different frequencies [52–54]. An audio signal filtered by the bottom convolutional layers of a one-dimensional CNN can be compared with $\mathbf{P}$. At this point, $\mathbf{P}$ is almost equivalent to the output of the bottom convolutional layers of a one-dimensional CNN. Figure 4 illustrates the flows of the inter-floor noise classification using the two-dimensional CNN (Figure 4a) and one-dimensional CNN (Figure 4b).

Figure 4. A flow of the inter-floor noise classification. (a) A signal from a detected onset position (Δ) s is converted to a log-scaled Mel-spectrogram P . P is provided to the three input channel of the feature extractor $f_{\Theta}(\cdot)$ then pseudo probabilities of categories in a label space for a given input $\hat{y}$ are predicted. (b) The one-dimensional CNN computes one-dimensional convolution directly on s and summarizes the signal.

Two-dimensional CNN-based classification starts with (1) conversion of a signal to a log-scaled Mel-spectrogram P , (2) convolutional layers and fully connected layers composing VGG16 as a feature extractor, (3) an adaptation layer (fc), and finishes with (4) classification using a softmax function. These were implemented with TensorFlow [55]. An inter-floor noise signal is converted to a log-scaled Mel-spectrogram $P \in \mathbb{R}^{H \times W}$ through the following steps, where the height H and width W are both defined as 224 by the input size of VGG16. This conversion was implemented with librosa [56]. A signal $s \in \mathbb{R}^{l_s}$ containing an inter-floor noise event is extracted from an audio clip in the dataset, where l_s represents f_s times the signal length t . The event start position in s is detected using the onset detection described in Section 3.1. Several $t = \{0.152, 0.501, 1.00, 1.50, 2.00, 3.00\}$ s are tested to identify the effect of t , where it can be considered a hyperparameter. The amplitudes of s are rescaled to values between $[-1, 1]$. Short time samples with size of $l = 1024$ from s are windowed, where the window is an $N = 2^{13}$ -point Hanning window. l sets the spectral line resolution as approximately 40 Hz for the given f_s . The windowed short time sample x_w , ($w = 0, \dots, W - 1$) is converted to a spectral power using a discrete Fourier transform

$$X_w(k) = |\text{DFT}[x_w(n)]|^2 = \left| \sum_{n=0}^{N-1} x_w(n) e^{-j2\pi nk/N} \right|^2, \quad k = 0, 1, \dots, N. \quad (6)$$

The start position of the next short time sample is determined by the hop size, $h = \{30, 99, 197, 296, 394, 591\}$ sample, to achieve W of P for the given t . A block of the windowed short time samples $x = [x_0 \mid \dots \mid x_{W-1}] \in \mathbb{R}^{N \times W}$ is converted to a power spectrogram X by Equation (6). Then, it is converted to a Mel-spectrogram

$$M = FX, \quad (7)$$

where $\mathbf{F}$ is a Mel-filterbank, which changes the scale of frequency to the Mel-scale. The maximum frequency of the filterbank is set as 5 kHz, because most of the signals in the dataset exist below this frequency, and other inter-floor noise studies deal with frequency ranges below 5 kHz [57,58]. The scale of the entries of $\mathbf{M}$ is also rescaled to obtain a log-scaled Mel-spectrogram, as follows:

$$\mathbf{P} = 10 \log(\mathbf{M} / \max(\mathbf{M}_{i,j})). \quad (8)$$

VGG16, the two-dimensional convolutional layers and fully connected layers employed in this work, was originally designed for image-recognition tasks. It has three input channels for the reception of a batch of color images, followed by 13 convolutional layers and three fully connected layers. Because a single two-dimensional feature is obtained via conversion of an inter-floor noise event to $\mathbf{P}$, the feature is given to all channels to train the weights Θ of the CNN and to test a trained CNN. As described in Section 2, the datasets obtained in this study are sparse compared to those adopted in other data-driven approaches. The Θ was initialized with weights pre-trained on a large-scale dataset (ImageNet) [59] to mitigate the problems originating from data sparsity. This can be viewed as transfer learning [60], which has already demonstrated its effectiveness in the image domain [61], as well as for inter-floor noise classification [9,10]. The image and the sound representations of the inter-floor noise data are considered different. However, the inter-floor noise events are converted to image-like representations $\mathbf{P}$ and used. Although the datasets are from different domains, low-level notions such as edges and shapes can be shared to learn distribution of a new task [48]. ImageNet is a large-scale dataset and adopted in transfer learning studies. Knowledge obtained from a large dataset has a higher chance to have sharable low-level notions than that via learning from scratch on a sparse dataset. In addition, it can prevent the over-fitting and contribute to the generalization. Use of low-level notions from a totally different domain knowledge and their performance improvements were presented [61,62]. In addition, use of a large-scale dataset also studied in low-shot learning [63]. This learns the metric using a large dataset and test it against unseen data. The output size of VGG16 is already given as 1000 by the source task, which is classification of ImageNet. It needs to be reduced to the size of the label space of a target $\text{Dim}(\mathcal{Y}_T)$. This was realized by adopting an additional fully connected layer, called the adaptation layer, with a reduced number of nodes $n = \text{Dim}(\mathcal{Y}_T)$. The weights between the last layer of VGG16 and the adaptation layer $\theta \in \mathbb{R}^{1000 \times n}$ were initialized as random numbers following a normal distribution with a standard deviation of 0.01 [50,51,64]. The bias of the adaptation layer is initialized as 1 [50]. The pseudo-probability of n categories $y_i \in \mathcal{Y}_T$, ($i = 1, \dots, n$) for a given input $\mathbf{P}$ is represented as $\hat{\mathbf{y}} \in \mathbb{R}^n$. It is obtained by inputting the output from the adaptation layer $\mathbf{o} \in \mathbb{R}^n$ into a softmax function

$$\sigma(\mathbf{o})_i = \frac{\exp(\mathbf{o}_i)}{\sum_{j=1}^n \exp(\mathbf{o}_j)}, \quad i = 1, \dots, n. \quad (9)$$

The classification into a category is

$$y = \underset{i}{\operatorname{argmax}} \sigma(\mathbf{o})_i. \quad (10)$$

A one-dimensional CNN computes one-dimensional convolution directly on the raw waveform and does not require feature engineering, such as converting audio signals to image-like features. SoundNet [65], as illustrated in Figure 4b, is adapted as a feature extractor after adding two adaptation layers to the top of the -1 convolutional layer (conv7) replacing the top convolutional layer (conv8). Because SoundNet is fully convolutional and summarizes raw waveforms via one-dimensional convolution and max-pooling, its waveforms are reduced to single values after passing through a few convolutional layers. Moreover, waveforms with different time lengths are summarized as different lengths of the outputs of the convolutional layers. Consequently, the number of weights of the adaptation

layers depends on the input size and this hinders the accurate identification of the effect of t . Therefore, $\mathbf{s}$ with $t = \{0.152, 0.501, 1.00, 1.50, 2.00, 3.00\}$ s was zero-padded to obtain an equal $t = 3.00$ s, and its amplitude was rescaled to values between $[-1, 1]$, which is filtered by the convolutional layers, and reaches an output size of 5120. The output from the convolutional layers is summarized by 5120×1024 (fc1) and $1024 \times n$ (fc2) weights. The Θ in SoundNet is initialized with weights pre-trained on a large-scale dataset (two million videos [65]) for finding a relationship between audio inputs and their corresponding image objects. The classification of inter-floor noise to source type or position categories is carried out using Equations (9) and (10).

3.3. Network Training

The Θ was trained by minimizing the cross-entropy loss and L_2 -regularization of the weights of the adaptation layers θ ,

$$\mathcal{L}_{CE} = - \sum_{i=1}^n \mathbf{y}_i \log \hat{\mathbf{y}}_i + \lambda \|\theta\|_2^2, \quad (11)$$

on a given training/validation dataset using the mini-batch gradient descent with batch size of 64, where $\mathbf{y}_i$ represents a one-hot-encoded label of a given $\mathbf{P}$. λ is a strength of the L_2 -regularization for θ to avoid overfitting.

The λ and learning rate η were selected using the random search method, which determines the optimal hyperparameters from the randomly sampled values [66]. Each hyperparameter follows the uniform distribution on the log-space in range of $[10^{-4}, 10^2]$. This search method reduces the effort for hyperparameter searching compared to the grid search method [67] and returns nearly optimal values. In this study, an optimal parameter pair of a given target domain $\mathcal{D}_T = \{\mathbf{P}_{i_k} \in \mathcal{X}_T, \mathcal{Y}_T\}$ is obtained via five-fold cross-validation, where i and k denote category and the number of data in the corresponding category, respectively. This can be realized as follows. (1) Fifty hyperparameter pairs are generated. (2) A hyperparameter pair with the maximum mean validation accuracy on five training/validation-folds is selected after 10 epochs of training. (3) Five predictive functions $f_{\Theta_j^*}(\cdot)$, ($j = 1, \dots, 5$) are obtained through 50 epochs of training on the five training/validation-folds, where Θ_j^* represents the weights of the whole network with the highest validation accuracy against the j th training/validation-fold.

3.4. Inter-Floor Noise Source Type Classification and Localization Tasks

Several source type classification and localization tasks are prepared to verify the CNN-based inter-floor noise classification and identify its limitations. These tasks verify the method via source type classification and localization in two apartment buildings, and knowledge transfer between two actual apartment buildings. The prepared tasks are presented in Tables 1 and 2 and explained in Sections 3.4.1–3.4.3. Table 1 presents tasks and datasets for training/validation and testing predictive functions. The notation of the task names in the first column is represented as $\mathcal{T}_{\text{task type, training/validation dataset}}$ or $\mathcal{T}_{\text{task type, test dataset|training/validation dataset}}$. Table 2 presents tasks prepared for verifying the knowledge transfer between the two apartment buildings. These tasks test the source type classifiers and locators using the inter-floor noise obtained from the other apartment building. The notation of the task name is represented as $\mathcal{T}_{\text{task type, test dataset|training/validation dataset}}$.

Table 1. Summary of the source type classification and localization tasks described in Sections 3.4.1 and 3.4.2 at the single apartment APT I and APT II.

Task	Task Type	Dataset	
		Training/Validation	Test
$\mathcal{T}_{t,APT 1}$ $\mathcal{T}_{t,APT 1' APT 1}$	Type	$\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{t,APT 1}$	None $\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{t,APT 1'}$
$\mathcal{T}_{t,APT 2}$ $\mathcal{T}_{t,APT 2' APT 2}$		$\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{t,APT 2}$	None $\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{t,APT 2'}$
$\mathcal{T}_{p,APT 1}$ $\mathcal{T}_{p,APT 1' APT 1}$	Position	$\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{p,APT 1}$	None $\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{p,APT 1'}$
$\mathcal{T}_{p,APT 2}$ $\mathcal{T}_{p,APT 2' APT 2}$		$\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{p,APT 2}$	None $\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{p,APT 2'}$

Table 2. Summary of knowledge transfer tasks described in Section 3.4.3.

Task	Task Type	Dataset	
		Training/Validation	Test
$\mathcal{T}_{t,APT II APT 1}$ $\mathcal{T}_{t,APT I APT 2}$	Type	$\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{t,APT 1}$ $\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{t,APT 2}$	$\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{t,APT II}$ $\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{t,APT I}$
$\mathcal{T}_{p,APT II APT 1}$ $\mathcal{T}_{p,APT I APT 2}$		$\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{p,APT 1}$ $\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{p,APT 2}$	$\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{p,APT II}$ $\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{p,APT I}$

3.4.1. Source Type Classification in a Single Apartment Building

The source type classification tasks summarized in Table 1 are described in the following (a)–(c). These tasks are prepared using the datasets from the two apartment buildings, APT I and APT II, to verify the CNN-based source type classification of inter-floor noise in a single apartment building.

- $\mathcal{T}_{t,APT 1}$. This task cross-validates the source type classification with the inter-floor noise on the floors above/below in APT I. This is realized by finding predictive functions $f_{\Theta_{t,APT 1}^j}(\cdot)$ with a label space $\mathcal{Y}_{t,APT 1}$ against five-folds ($j = 1, \dots, 5$) of labeled training/validation data pairs $\{\mathbf{P}_{i_k}, y_i\}$ from $\mathcal{D}_{APT 1}$.
- $\mathcal{T}_{t,APT 1'|APT 1}$. This task verifies the source type classification against the inter-floor noise generated at unlearned positions in the same apartment building, APT I. $f_{\Theta_{t,APT 1}^j}(\cdot)$ obtained in $\mathcal{T}_{t,APT 1}$ is tested against the test data pairs $\{\mathbf{P}_{i_k}, y_i\}$ from $\mathcal{D}_{APT 1'}$.
- $\mathcal{T}_{t,APT 2}$ and $\mathcal{T}_{t,APT 2'|APT 2}$. These tasks verify the same things against the noise samples on the floors above/below in APT II. $f_{\Theta_{t,APT 2}^j}(\cdot)$, ($j = 1, \dots, 5$) with $\mathcal{Y}_{t,APT 2}$ are obtained and tested against $\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{t,APT 2'}$ from $\mathcal{D}_{APT 2'}$.

3.4.2. Localization in a Single Apartment Building

Localization tasks presented in Table 1 are described in the following (a)–(c). These tasks are prepared to verify the localization of inter-floor noise in a single building (APT I and APT II). In addition, localization of inter-floor noise from the same XY position but transmitted through different floor sections (different Z position) is verified.

- $\mathcal{T}_{p,APT 1}$. This task cross-validates locators against the inter-floor noise on the floors above/below in APT I. This is realized by finding $f_{\Theta_{p,APT 1}^j}(\cdot)$ with $\mathcal{Y}_{p,APT 1}$ on five-folds ($j = 1, \dots, 5$) of labeled training/validation data pairs $\{\mathbf{P}_{i_k}, y_i\}, y_i \in \mathcal{Y}_{t,APT 1}$ from $\mathcal{D}_{APT 1}$.
- $\mathcal{T}_{p,APT 1'|APT 1}$. $f_{\Theta_{p,APT 1}^j}(\cdot)$ obtained in $\mathcal{T}_{p,APT 1}$ are tested against the inter-floor noises generated at the unlearned positions in the same apartment building, APT I. The pre-

- dictive functions are tested against the test data pairs $\{\mathbf{P}_{i_k}, y_i\}$, $y_i \in \mathcal{Y}_{t, \text{APT } 1'}$ from $\mathcal{D}_{\text{APT } 1'}$.
- (c) $\mathcal{T}_{p, \text{APT } 2}$ and $\mathcal{T}_{p, \text{APT } 2' | \text{APT } 2}$. They verify locators using the same approach for $\mathcal{T}_{p, \text{APT } 1}$ and $\mathcal{T}_{p, \text{APT } 1' | \text{APT } 1}$ with the inter-floor noise obtained from APT II.

3.4.3. Knowledge Transfer between the Apartment Buildings

If the trained source type and localization knowledge can be reused for inter-floor noise identification tasks, without required training or training under data sparsity on samples from other similar apartment building, then this method can be used widely. The knowledge transfer tasks presented in Table 2 are described in the following (a) and (b).

- (a) $\mathcal{T}_{t, \text{APT II} | \text{APT I}}$ and $\mathcal{T}_{t, \text{APT I} | \text{APT II}}$. These tasks test $f_{\Theta_{t, \text{APT I}}}^j(\cdot)$ and $f_{\Theta_{t, \text{APT II}}}^j(\cdot)$ against data pairs from $\mathcal{D}_{\text{APT II}}$ and $\mathcal{D}_{\text{APT I}}$, respectively.
- (b) $\mathcal{T}_{p, \text{APT II} | \text{APT I}}$ and $\mathcal{T}_{p, \text{APT I} | \text{APT II}}$. These tasks test localization knowledge transfer. The XY positions relative to the receiver position of the data points in APT I and APT II are different. Therefore, the position label space of APT I and APT II are considered different. Hence, $f_{\Theta_{p, \text{APT I}}}^j(\cdot)$ and $f_{\Theta_{p, \text{APT II}}}^j(\cdot)$ are tested against data pairs from $\mathcal{D}_{\text{APT II}}$ and $\mathcal{D}_{\text{APT I}}$, respectively. Furthermore, the localized positions are rearranged to their corresponding floors.

4. Performance Evaluation

F1 score in Equation (12) is implemented with Scikit-learn [68] and adopted to measure the performance of the predictive functions. TP, FP, and FN represent true positive, false positive, and false negative, respectively. This metric evaluates the classification results considering the imbalance of the inter-floor noise dataset.

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}, \quad (12)$$

$$\text{where Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \text{ and } \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}.$$

4.1. Source Type Classification Results in a Single Apartment Building

Table 3 presents F1 scores of the source type classification results with t variation. The F1 scores with HH and HD set to the same category are placed in parentheses to distinguish them from the results of the original task. As exhibited by the differences between the F1 scores in two representations, $f_{\Theta_{t, \text{APT I}}}^j(\cdot)$ and $f_{\Theta_{t, \text{APT II}}}^j(\cdot)$ confuse HH and HD for all tasks. Figure 5 shows the time–frequency representations of HD and HH at the campus building (Figure 5a,d), APT I (Figure 5b,e), and APT II (Figure 5c,f). HD at the campus building shows repeated impact noise patterns (peaks) induced by bouncing of the hammer when it was dropped on terrazzo tile flooring. However, HD obtained from both apartment buildings was generated with a hammer on the slabs covered with vinyl floorings, which prevented the hammer from bouncing and mitigated the peak patterns. Hence, the source type classification results of assuming HH and HD as the same category need to be additionally considered. Figure 6a presents trend of the F1 scores of the source type classification results with t variation. This plot visualizes the results of the test tasks ($\mathcal{T}_{t, \text{APT } 1' | \text{APT } 1}$ and $\mathcal{T}_{t, \text{APT } 2' | \text{APT } 2}$). The F1 scores of the source type predictive functions on inter-floor noise from the unlearned positions, are lower than those of the cross-validation tasks ($\mathcal{T}_{t, \text{APT } 1}$ and $\mathcal{T}_{t, \text{APT } 2}$). SoundNet-based classifiers under-performed for source type classification tasks, as well as exhibited significantly low F1 scores for s at $t = 0.152$ and 0.501 s. Additionally, t influences the similar variations of F1 scores related to the VGG16-based classification.

In summary, the CNN-based approach demonstrated the feasibility and generalizability of the source type classification to inter-floor noise from actual reinforced concrete apartment buildings. The adapted two-dimensional CNN exhibited marginally better

performance than that of the adapted one-dimensional CNN. This may be explained by the sharable low-level notions in the pre-trained knowledge.

Figure 5. Time–frequency representations of inter-floor noise signals. (a,d) show HD and HH at the campus building. (b,e) show HD and HH in APT I. (c,f) show HD and HH in APT II.

Table 3. F1 scores of source type classification tasks with t variation. The values in the parentheses are F1 scores when HH and HD are assumed to be the same category.

Task	CNN	t (s)					
		0.152	0.501	1.00	1.50	2.00	3.00
$\mathcal{T}_{t,APT 1}$	VGG16	0.9346 (0.9377)	0.9663 (0.9708)	0.9663 (0.9742)	0.9719 (0.9764)	0.9731 (0.9798)	0.9662 (0.9730)
	SoundNet	0.3016 (0.4162)	0.4102 (0.4913)	0.9497 (0.9553)	0.9555 (0.9634)	0.9545 (0.9601)	0.9172 (0.9240)
$\mathcal{T}_{t,APT 1' APT 1}$	VGG16	0.6801 (0.9606)	0.8235 (0.9645)	0.8019 (0.9552)	0.7873 (0.9220)	0.8303 (0.9563)	0.4616 (0.5128)
	SoundNet	0.2286 (0.3536)	0.3339 (0.4094)	0.6862 (0.8229)	0.7494 (0.8132)	0.7324 (0.8559)	0.6976 (0.7688)
$\mathcal{T}_{t,APT 2}$	VGG16	0.9513 (0.9917)	0.9582 (0.9938)	0.9536 (0.9948)	0.9541 (0.9938)	0.9551 (0.9953)	0.9456 (0.9922)
	SoundNet	0.3491 (0.5523)	0.3789 (0.5266)	0.9487 (0.9886)	0.9507 (0.9876)	0.9501 (0.9876)	0.9378 (0.9844)
$\mathcal{T}_{t,APT 2' APT 2}$	VGG16	0.7038 (0.9664)	0.7262 (0.9898)	0.8553 (0.9896)	0.7910 (0.9907)	0.7991 (0.9875)	0.9102 (0.9900)
	SoundNet	0.2965 (0.4530)	0.3204 (0.4686)	0.8079 (0.9106)	0.8616 (0.9164)	0.8064 (0.9145)	0.7324 (0.9249)

In addition, appropriate selection of t influences the performance with improved F1 scores of cross-validation, as well as test tasks against the inter-floor noise transmitted through the unlearned floor sections.

4.2. Localization Results in a Single Apartment Building

The first eight rows of Table 4 present F1 scores of the localization results via five-fold cross-validation, $\mathcal{T}_{p,APT 1}$ and $\mathcal{T}_{p,APT 2}$, with t variation. The underlined values represent F1 scores of the floor classification via rearrangement of the localization results to their corresponding floors. Because 95.7% of the actual inter-floor noise complaints were identified as noise from floors above/below [4], these floor classifications are considered to be the main

interest in real application. Figure 6b,c present trend of the localization results and those of the floor classification tasks with t variation, respectively.

The remaining eight rows present F1 scores of test results via $\mathcal{T}_{p,APT 1'|APT 1}$ and $\mathcal{T}_{p,APT 2'|APT 2}$. $\mathcal{T}_{p,APT 1'|APT 1}$ tests $f_{\Theta_{p,APT 1}^j}(\cdot)$ on $\{\mathbf{P}_{i_k}, y_i\}$, $y_i \in \mathcal{Y}_{p,APT 1'}$. However, there exists a label-space difference between $f_{\Theta_{p,APT 1}^j}(\cdot)$ (i.e., $\mathcal{Y}_{p,APT 1}$) and $\mathcal{Y}_{p,APT 1'}$. Accordingly, 1-C-a, 1-D-a, and 1-E-a cannot be considered any of the categories learned by $f_{\Theta_{p,APT 1}^j}(\cdot)$ because their XY positions relative to the receiver are different from those learned by $f_{\Theta_{p,APT 1}^j}(\cdot)$. Therefore, the localization test results are evaluated separately, as follows.

- The position labels 1-A'-a and 1-B'-a are considered 1-A-a and 1-B-a, respectively. These realize the localization of inter-floor noise transmitted through the unlearned floor section (4 F $\rightarrow$ 3 F in APT I).
- Localization results of 1-C-a, 1-D-a and 1-E-a are squeezed into $y_i \in \mathcal{Y}_{p,APT 1}$ and approximated to floor classification because their XY positions cannot be mapped to $y_i \in \mathcal{Y}_{p,APT 1}$ directly.

If the noise sources with the same XY positions (e.g., 2-A-a and 2-A'-a) are assumed to be the same category, $\{\mathbf{P}_{i_k}, y_i\}$, $y_i \in \mathcal{Y}_{p,APT 2'}$ for $\mathcal{T}_{p,APT 2'|APT 2}$ can be mapped to $y_i \in \mathcal{Y}_{p,APT 2}$ by $f_{\Theta_{p,APT 2}^j}(\cdot)$. The locators presented a clear limitation on localization of inter-floor noise transmitted through the two unlearned floor sections, 4 F $\rightarrow$ 3 F in APT I and 5 F $\rightarrow$ 4 F in APT II, respectively. On the other hand, the locator showed the feasibility on floor classification of inter-floor noise transmitted through the unlearned floor sections.

Table 4. F1 scores of the position classification tasks with t variation. The underlined values are the floor classification F1-scores obtained by rearrangement of the results.

Task	CNN	t (s)					
		0.152	0.501	1.00	1.50	2.00	3.00
$\mathcal{T}_{p,APT 1}$	VGG16	0.9047	0.9381	0.9438	0.9516	0.9574	0.9496
		<u>0.9607</u>	<u>0.9899</u>	<u>0.9910</u>	<u>1.000</u>	<u>0.9955</u>	<u>0.9944</u>
	SoundNet	0.7526	0.6954	0.7338	0.9426	0.9323	0.9607
		<u>0.9786</u>	<u>0.9273</u>	<u>0.9764</u>	<u>0.9910</u>	<u>0.9865</u>	<u>0.9899</u>
$\mathcal{T}_{p,APT 2}$	VGG16	0.8344	0.8785	0.8968	0.9079	0.9272	0.9328
		<u>0.9333</u>	<u>0.9557</u>	<u>0.9682</u>	<u>0.9813</u>	<u>0.9786</u>	<u>0.9807</u>
	SoundNet	0.4413	0.5150	0.8787	0.8752	0.9017	0.9337
		<u>0.8512</u>	<u>0.9318</u>	<u>0.9646</u>	<u>0.9641</u>	<u>0.9750</u>	<u>0.9880</u>
$\mathcal{T}_{p,APT 1' APT 1}$	VGG16	0.6794	0.7401	0.7805	0.5960	0.7032	0.5596
		<u>0.9709</u>	<u>0.9651</u>	<u>0.9796</u>	<u>0.9818</u>	<u>0.9834</u>	<u>0.9793</u>
	SoundNet	0.1118	0.3461	0.4336	0.5951	0.6967	0.5093
		<u>0.2038</u>	<u>0.8627</u>	<u>0.9564</u>	<u>0.9087</u>	<u>0.9276</u>	<u>0.9333</u>
$\mathcal{T}_{p,APT 2' APT 2}$	VGG16	0.2857	0.3454	0.3211	0.3470	0.2818	0.3475
		<u>0.9570</u>	<u>0.9635</u>	<u>0.9811</u>	<u>0.9775</u>	<u>0.9704</u>	<u>0.9752</u>
	SoundNet	0.1554	0.1855	0.3506	0.3557	0.3837	0.3356
		<u>0.6381</u>	<u>0.7368</u>	<u>0.9088</u>	<u>0.8348</u>	<u>0.9358</u>	<u>0.9059</u>

In summary, the CNN-based locator demonstrated the feasibility and generalizability of floor classification of inter-floor noise generated on the floors above/below, and those transmitted through the learned/unlearned floor sections in actual apartment buildings. It can contribute to minimize human effort for data gathering pertaining to the floor classification problem. However, localization of inter-floor noise was available for that transmitted through the learned floor sections alone. The adapted one-dimensional CNN-based locator with $t = 0.152$ and 0.501 s exhibited significantly low F1 scores.

4.3. Results of Knowledge Transfer between the Apartment Buildings

Table 5 presents evaluated performance of the knowledge transfer tasks with t variation, as explained in Section 3.4.3. The first eight rows present results of the source type knowledge transfer tasks $\mathcal{T}_{t,APT II|APT 1}$ and $\mathcal{T}_{t,APT I|APT 2}$. Their performance was evaluated with F1 scores, the values in parentheses are F1 scores when HH and HD are assumed to be the same category. The source type knowledge transfer results showed lower F1 scores than those of the cross-validation and test tasks. This performance degradation may originate from the domain difference. The underlined values in the remaining rows present F1 scores of the localization knowledge transfer results of $\mathcal{T}_{p,APT II|APT 1}$ and $\mathcal{T}_{p,APT I|APT 2}$. They were obtained by rearranging the localization results to their corresponding floors. As illustrated in Figure 6d, these F1 scores degraded as t increased. These may be explained by the difference of the reverberation characteristic between the inter-floor noise obtained in the two apartment buildings. Therefore, when target environment (apartment building) is changed, using short and early part of inter-floor noise is effective for localization with the source domain knowledge. Although the F1 scores of the results are lower than those of the cross-validation and the test tasks, they are considered to be meaningful, because they are better than the chance level, and demonstrate the feasibility of source type and localization knowledge transfer between the actual apartment buildings. The performance degradation is common for the one-dimensional CNN-based locator with $t = 0.152$ and 0.501 s.

The feasibility of the knowledge transfer can be considered, such that a CNN can extract the generalized feature representations of source types and positions from inter-floor noises. Inter-floor noise filtered by learned one-dimensional CNNs with $t = 1.00$ s for source type classification and localization were embedded in a two-dimensional space using a dimension reduction algorithm t-stochastic neighbor embedding (t-SNE) [69] for visualization of the generalized features. Because the dimension of fc2 depends on the number of categories (n), output from fc1 were adopted in this case. Figures 7 and 8 present the t-SNE of the inter-floor noises filtered by the CNNs. Figure 7a,b illustrate the t-SNE of source type features in $\mathcal{T}_{t,APT II|APT 1}$ and $\mathcal{T}_{t,APT I|APT 2}$, respectively. The t-SNE of the same source types within the same apartment buildings are categorized into groups. In particular, the t-SNE of the same source types in different apartment buildings are clustered around the groups. This demonstrates that the CNN-based feature extraction can build generalized source type representations of the inter-floor noises. Figure 8 illustrates the t-SNE of floor features in $\mathcal{T}_{p,APT II|APT 1}$ and $\mathcal{T}_{p,APT I|APT 2}$. The t-SNE of fc1 are categorized into groups of above and below floors. However, the groups are not well clustered as the t-SNE of the source type features. It can be inferred that although the feature extraction of a generalized floor knowledge using CNN is feasible, but it is not significantly effective. The results in Table 5 also exhibit the corresponding values.

Table 5. Inter-floor noise classification knowledge transfer results with t variation. The values are F1 score of the source type knowledge transfer results. The values in the parentheses are the F1 scores when the HH and HD categories are assumed to be the same category. The underlined values are rearranged localization knowledge transfer results to their corresponding floors.

Task	CNN	t (s)					
		0.152	0.501	1.00	1.50	2.00	3.00
$\mathcal{T}_{t, \text{APT II} \text{APT I}}$	VGG16	0.5778 (0.8273)	0.6244 (0.8520)	0.5865 (0.8178)	0.5952 (0.8356)	0.6617 (0.8800)	0.2238 (0.3044)
	SoundNet	0.2557 (0.3727)	0.2318 (0.3274)	0.5789 (0.7838)	0.5636 (0.7661)	0.5421 (0.7848)	0.4995 (0.7396)
$\mathcal{T}_{t, \text{APT I} \text{APT II}}$	VGG16	0.5296 (0.7285)	0.7522 (0.8046)	0.7482 (0.7957)	0.7682 (0.7970)	0.6122 (0.7656)	0.7684 (0.8093)
	SoundNet	0.2048 (0.3231)	0.2196 (0.3385)	0.6009 (0.6537)	0.5196 (0.5977)	0.5691 (0.6481)	0.4876 (0.5674)
$\mathcal{T}_{p, \text{APT II} \text{APT I}}$	VGG16	<u>0.8037</u>	<u>0.7842</u>	<u>0.7384</u>	<u>0.6786</u>	<u>0.6976</u>	<u>0.5678</u>
	SoundNet	<u>0.4528</u>	<u>0.4741</u>	<u>0.5747</u>	<u>0.6941</u>	<u>0.7943</u>	<u>0.6443</u>
$\mathcal{T}_{p, \text{APT I} \text{APT II}}$	VGG16	<u>0.8903</u>	<u>0.8309</u>	<u>0.7654</u>	<u>0.7057</u>	<u>0.6850</u>	<u>0.6006</u>
	SoundNet	<u>0.2171</u>	<u>0.2969</u>	<u>0.7650</u>	<u>0.7513</u>	<u>0.7094</u>	<u>0.7290</u>

Figure 6. Results of inter-floor classification and classification knowledge-transfer tasks evaluated using F1 score. HD and HH were considered the same category for source type classification results. (a) Cross-validation and test results of the source type classification tasks in Table 3. (b) Cross-validation and test results of localization tasks in Table 4. (c) Rearranged localization results to their corresponding floor. (d) Results of the knowledge transfer tasks in Table 5.

Figure 7. t-SNE of source type features filtered by (a) SoundNet $f_{\Theta_{t,APT 1}^j}(\cdot)$ in $\mathcal{T}_{t,APT II|APT 1}$ and (b) SoundNet $f_{\Theta_{t,APT 2}^j}(\cdot)$ in $\mathcal{T}_{t,APT I|APT 2}$. The symbols in purple and those in yellow represent t-SNE of learned and unlearned signals, respectively, for a single selected source type. In addition, those in gray represent the other source types.

Figure 8. The symbols in purple and those in yellow represent t-SNE of learned and unlearned signals, respectively, for a single selected floor among floor above and below. In addition, those in gray represent the other floor. (a,b), respectively, colorize signals from the floor above and below in $\mathcal{T}_{p,APT II|APT 1}$. (c,d), respectively, colorize signals from the floor above and below in $\mathcal{T}_{p,APT I|APT 2}$.

4.4. Input Signal Length Selection

It is recommended to fix t for implementation of the method for real-world application. The SoundNet-based predictive functions exhibited a lower performance in all inter-floor noise identification except in $\mathcal{T}_{p,APT 2}$, as well as significantly low F1 scores for s at $t = 0.152$ and 0.501 s. Therefore, VGG16-based predictive function is considered alone in this discussion. The source type classifiers and locators with $t = 1.50$ and 2.00 s exhibited the best F1 scores most frequently. Additionally, the locators with $t = 0.501$ s showed clear effectiveness for localization knowledge transfer.

5. Conclusions and Future Study

In this study, the generalizability of the CNN-based supervised learning method for source type classification and localization of inter-floor noise was demonstrated via several designed tasks. Furthermore, the feasibility of source type and localization knowledge

transfer between apartment buildings was demonstrated. These were demonstrated using inter-floor noise datasets obtained from two actual apartment buildings.

The source type classifiers and locators consist of CNN-based feature extractors followed by an adaptation layers and a softmax function. Pre-trained weights on largely annotated image datasets were used for initialization of the feature extractors. Noise events in the signals were detected using the HOS-based algorithm and inputted to one-dimensional (SoundNet) and two-dimensional (VGG16) CNNs. The signal length in time t was selected empirically.

The source type classifiers and locators were verified against several tasks, which are five-fold cross-validation on inter-floor noise transmitted through a learned floor section in individual two apartment buildings, test on that transmitted through an unlearned floor section in the same apartment buildings, and test on that obtained from unseen apartment buildings. The performance of the method for each task was evaluated using the F1 score. VGG16-based source type classifier and locator performed better than those of SoundNet-based. For cross-validation of the source type classification with $t = 2.00$ s, VGG16-based source type classifier showed F1 scores of 0.9731 and 0.9551 for APT I and APT II dataset, respectively. If HD and HH are considered the same category, these scores were improved to 0.9798 and 0.9953. For test of the prepared source type classification tasks using inter-floor noises transmitted from the unlearned source positions, F1 scores of source type classification results by the same classifiers dropped to 0.8303 and 0.7991, respectively. Furthermore, if HD and HH are considered the same category, the scores reached 0.9563 and 0.9875. For cross-validation of the localization, VGG16-based locator with $t = 2.00$ s showed F1 scores of 0.9574 and 0.9272 for APT I and APT II dataset, respectively. Rearranging these results to their corresponding floors modified F1 scores to 0.9955 and 0.9786, respectively. However, the method showed limitation to test tasks using noise signals transmitted through the unlearned floor sections. However, rearranged results to corresponding floors classification dropped less than 2% from the those of cross-validation tasks. These results present the generalizability of the identification method with a single sensor in an actual multi-storey reinforced concrete building and the feasibility of the knowledge transfer between similar buildings.

In conclusion, results in this study contribute to identify inter-floor noise in real multi-storey apartment buildings and other single sensor-based approaches. Future studies should focus on resolving the limitations of the method by using the geometrical information of the buildings for localization of signals at unlearned positions or domain adaptation techniques for better knowledge transfer. In addition, although the CNN-based approaches showed their generalizability, the datasets are only suitable for the designed tasks. To deal with tasks with a high degree of freedom, the data are expected to be updated considering many different scenarios for better generalization, e.g., change of the receiver's position and orientation.

Author Contributions: Conceptualization, W.S. and H.Y.; methodology, H.C., W.S. and H.Y.; software, H.C.; validation, W.S. and H.Y.; data curation, H.C., W.S. and H.Y.; writing, H.C., W.S. and H.Y.; supervision, W.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Research Foundation of Korea (NRF) with a grant funded by the Korea government (MSIT) (No. NRF-2017R1E1A2A01078766) and (No. NRF-2019R1F1A1058794) and supported by the Agency for Defense Development in Korea under Contract No. UD190005DD.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Acknowledgments: The authors thank Sangkyum An and Minseuk Park for providing the experimental sites.

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

References

1. Jeon, J.Y.; Ryu, J.K.; Lee, P.J. A quantification model of overall dissatisfaction with indoor noise environment in residential buildings. *Appl. Acoust.* **2010**, *71*, 914–921. [CrossRef]
2. Jeon, J.Y. Subjective evaluation of floor impact noise based on the model of ACF/IACF. *J. Sound Vib.* **2001**, *241*, 147–155. [CrossRef]
3. Maschke, C.; Niemann, H. Health effects of annoyance induced by neighbour noise. *Noise Control Eng. J.* **2007**, *55*, 348–356. [CrossRef]
4. Floor Noise Management Center. Monthly Report on Inter-Floor Noise Complaints. Available online: http://www.noiseinfo.or.kr/about/data_view.jsp?boardNo=199&keyfield=whole&keyword=&pg=2 (accessed on 19 January 2021).
5. Park, S.H.; Lee, P.J.; Yang, K.S.; Kim, K.W. Relationships between non-acoustic factors and subjective reactions to floor impact noise in apartment buildings. *J. Acoust. Soc. Am.* **2016**, *139*, 1158–1167. [CrossRef] [PubMed]
6. Korean Statistical Information Service. 2019 Housing Units by Type of Housing Units. Available online: https://kosis.kr/eng/statisticsList/statisticsListIndex.do?menuId=M_01_01&vwcd=MT_ETITLE&parmTabId=M_01_01&statId=1962005&themaId=#SelectStatsBoxDiv (accessed on 4 June 2021).
7. Korea Environment Corporation. Korea Environment Corporation Main Webpage. Available online: <https://www.keco.or.kr/en/main/index.do> (accessed on 19 January 2021).
8. Floor Noise Management Center. Inter-Floor Noise Complaints Received until the End of Year 2019. Available online: http://www.noiseinfo.or.kr/about/stats/counselServiceSttus_01.jsp (accessed on 19 January 2021).
9. Choi, H.; Lee, S.; Yang, H.; Seong, W. Classification of noise between floors in a building using pre-trained deep convolutional neural networks. In Proceedings of the 16th International Workshop on Acoustic Signal Enhancement (IWAENC), Tokyo, Japan, 17–20 September 2018; pp. 535–539.
10. Choi, H.; Yang, H.; Lee, S.; Seong, W. Classification of inter-floor noise type/position via convolutional neural network-based supervised learning. *Appl. Sci.* **2019**, *18*, 3735. [CrossRef]
11. Bahroun, R.; Michel, O.; Frassati, F.; Carmona, M.; Lacoume, J.L. New algorithm for footstep localization using seismic sensors in an indoor environment. *J. Sound Vib.* **2014**, *333*, 1046–1066. [CrossRef]
12. Poston, J.D.; Buehrer, R.M.; Tarazaga, P.A. Indoor footstep localization from structural dynamics instrumentation. *Mech. Syst. Signal Process.* **2017**, *88*, 224–239. [CrossRef]
13. Mirshekari, M.; Pan, S.; Fagert, J.; Schooler, E.M.; Zhang, P.; Noh, H.Y. Occupant localization using footstep-induced structural vibration. *Mech. Syst. Signal Process.* **2018**, *112*, 77–97. [CrossRef]
14. Barchiesi, D.; Giannoulis, D.; Stowell, D.; Plumbley, M.D. Acoustic scene classification: Classifying environments from the sounds they produce. *IEEE Signal Process. Mag.* **2015**, *32*, 16–34. [CrossRef]
15. Abeßer, J. A review of deep learning based methods for acoustic scene classification. *Appl. Sci.* **2020**, *10*, 2020. [CrossRef]
16. Sawhney, N.; Maes, P. Situational awareness from environmental sounds. *Proj. Rep. Pattie Maes* **1997**, 1–7.
17. Malkin, R.G.; Waibel, A. Single-channel indoor microphone localization. In Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), Philadelphia, PA, USA, 18–23 March 2005; pp. 1434–1438.
18. Aucouturier, J.J.; Defreville, B.; Pachet, F. The bag-of-frames approach to audio pattern recognition: A sufficient model for urban soundscapes but not for polyphonic music. *J. Acoust. Soc. Am.* **2007**, *122*, 881–891. [CrossRef]
19. Piczak, K.J. Environmental sound classification with convolutional neural networks. In Proceedings of the IEEE 25th International Workshop on Machine Learning for Signal Processing (MLSP), Boston, MA, USA, 17–20 September 2015; pp. 1–6.
20. Tobias, A. Acoustic-emission source location in two dimensions by an array of three sensors. *Non Destruct. Test* **1976**, *9*, 9–12. [CrossRef]
21. Ciampa, F.; Meo, M. Acoustic emission localization in complex dissipative anisotropic structures using a one-channel reciprocal time reversal method. *J. Acoust. Soc. Am.* **2011**, *130*, 168–175. [CrossRef] [PubMed]
22. Kundu, T.; Das, S.; Jata, K.V. Point of impact prediction in isotropic and anisotropic plates from the acoustic emission data. *J. Acoust. Soc. Am.* **2007**, *122*, 2057–2066. [CrossRef]
23. Ciampa, F.; Meo, M. Acoustic emission source localization and velocity determination of the fundamental mode A0 using wavelet analysis and a newton-based optimization technique. *Smart Mater. Struct.* **2010**, *19*, 045027. [CrossRef]
24. Goutaudier, D.; Gendre, D.; Kehr-Candille, V.; Ohayon, R. Single-sensor approach for impact localization and force reconstruction by using discriminating vibration modes. *Mech. Syst. Signal Process.* **2020**, *138*, 106534. [CrossRef]
25. Grabec, I.; Sachse, W. Application of an intelligent signal processing system to acoustic emission analysis. *J. Acoust. Soc. Am.* **1989**, *85*, 1226–1235. [CrossRef]
26. Kosel, T.; Grabec, I.; Mužič, P. Location of acoustic emission sources generated by air flow. *Ultrasonics* **2000**, *38*, 824–826. [CrossRef]
27. Ing, R.K.; Quieffin, N.; Catheline, S.; Fink, M. In solid localization of finger impacts using acoustic time-reversal process. *Appl. Phys. Lett.* **2005**, *87*, 204104. [CrossRef]
28. Ruiz, M.; Mujica, L.; Berjaga, X.; Rodellar, J. Partial least square/projection to latent structures (PLS) regression to estimate impact localization in structures. *Smart Mater. Struct.* **2013**, *22*, 025028. [CrossRef]
29. Sung, D.U.; Oh, J.H.; Kim, C.G.; Hong, C.S. Impact monitoring of smart composite laminates using neural network and wavelet analysis. *J. Intell. Mater. Syst. Struct.* **2000**, *11*, 180–190. [CrossRef]
30. Ebrahimkhanlou, A.; Salamone, A. Single-sensor acoustic emission source localization in plate-like structures using deep learning. *Aerospace* **2018**, *5*, 50. [CrossRef]

31. Parhizkar, R.; Dokmanić, I.; Vetterli, M. Single-channel indoor microphone localization. In Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 1434–1438.
32. Niu, H.; Gong, Z.; Ozanich, E.; Gerstoft, P.; Wang, H.; Li, Z. Deep-learning source localization using multi-frequency magnitude-only data. *J. Acoust. Soc. Am.* **2019**, *146*, 211–222. [\[CrossRef\]](#)
33. Komen, D.F.; Neilsen, T.B.; Howarth, K.; Knobels, D.P.; Dahl, P.H. Seabed and range estimation of impulsive time series using a convolutional neural network. *J. Acoust. Soc. Am.* **2020**, *147*, EL403–EL408. [\[CrossRef\]](#)
34. Poston, J.D. Toward tracking multiple building occupants by footstep vibrations. In Proceedings of the IEEE Global Conference on Signal and Information Processing (GlobalSIP), Anaheim, CA, USA, 26–28 November 2018; pp. 86–90.
35. Woolard, A.G. Supplementing Localization Algorithms for Indoor Footsteps. Ph.D. Thesis, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA, 7 July 2017.
36. The Ministry of Land, Infrastructure and Transport Korea. Statistics of Housing Construction (Construction Consent). Available online: http://kosis.kr/statHtml/statHtml.do?orgId=116&tblId=DT_MLTM_564&conn_path=I2 (accessed on 19 January 2021).
37. The Seoul Institute. Construction Consent. Available online: <http://data.si.re.kr/node/344> (accessed on 19 January 2021).
38. Song, Y.; Choi, G.-R. Flat column dry wall (FcDW) system design for apartment. *Mag. Korea Concr. Inst.* **2008**, *20*, 37–42.
39. Chosun Ilbo. Noise Characteristic of Apartment Buildings in the South Korea. Available online: http://realty.chosun.com/site/data/html_dir/2018/08/21/2018082102461.html (accessed on 19 January 2021).
40. Samsung Electronics. Galaxy S6. Available online: <https://www.samsung.com/global/galaxy/galaxys6/galaxy-s6> (accessed on 11 March 2020).
41. Allen, R.V. Automatic earthquake recognition and timing from single traces. *Bull. Seismol. Soc. Am.* **1978**, *68*, 1521–1532.
42. Kurz, J.H.; Grosse, C.U.; Reinhardt, H.W. Strategies for reliable automatic onset time picking of acoustic emissions and of ultrasound signals in concrete. *Ultrasonics* **2005**, *43*, 538–546. [\[CrossRef\]](#)
43. Allen, R.V. Automatic phase pickers: Their present use and future prospects. *Bull. Seismol. Soc. Am.* **1982**, *72*, S225–S242.
44. Hensman, J.; Mills, R.; Pierce, S.G.; Worden, K.; Eaton, M. Locating acoustic emission sources in complex structures using Gaussian processes. *Mech. Syst. Signal Process.* **2010**, *24*, 211–223. [\[CrossRef\]](#)
45. Chung, P.; Jost, M.L.; Böhme, J.F. Estimation of seismic-wave parameters and signal detection using maximum-likelihood methods. *Comput. Geosci.* **2001**, *27*, 147–156. [\[CrossRef\]](#)
46. Saragiotis, C.D.; Hadjileontiadis, L.J.; Panas, S.M. PAI-S/K: A robust automatic seismic P phase arrival identification scheme. *IEEE Trans. Geosci. Remote. Sens.* **2002**, *40*, 1395–1404. [\[CrossRef\]](#)
47. Saragiotis, C.D.; Hadjileontiadis, L.J.; Rekanos, I.T.; Panas, S.M. Automatic P phase picking using maximum kurtosis and/spl kappa/-statistics criteria. *IEEE Trans. Geosci. Remote. Sens. Lett.* **2004**, *1*, 147–151. [\[CrossRef\]](#)
48. Goodfellow, I.; Bengio, Y.; Courville, A. Representation learning. In *Deep learning*; Dietterich, T., Bishop, C., Heckerman, D., Jordan, M., Kearns, M., Eds.; The MIT Press: Cambridge, MA, USA; London, UK, 2017; pp. 330–372. ISBN 978-026-203-561-3.
49. 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, 26 June–1 July 2016; pp. 770–778.
50. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems (NIPS), Lake Tahoe, NV, USA, 3–6 December 2012; pp. 1097–1105.
51. Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv* **2014**, arXiv:1409.1556.
52. Dieleman, S.; Schrauwen, B. End-to-end learning for music audio. In Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), Florence, Italy, 4–9 May 2014; pp. 6964–6968.
53. Tokozume, Y.; Harada, T. Learning environmental sounds with end-to-end convolutional neural network. In Proceedings of the IEEE Conference on Acoustics, Speech and Signal Processing (ICASSP), New Orleans, LA, USA, 5–9 March 2017; pp. 2721–2725.
54. Lee, J.; Park, J.; Kim, K.L.; Nam, J. End-to-end deep convolutional neural networks using very small filters for music classification. *Appl. Sci.* **2018**, *8*, 150. [\[CrossRef\]](#)
55. Abadi, M.; Agarwal, A.; Barham, P.; Brevdo, E.; Chen, Z.; Citro, C.; Corrado, G.S.; Davis, A.; Dean, J.; Devin, M.; et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. Available online: <https://www.tensorflow.org> (accessed on 20 March 2020).
56. McFee, B.; Raffel, C.; Liang, D.; Ellis, D.P.; McVicar, M.; Battenberg, E.; Nieto, O. librosa: Audio and music signal analysis in python. In Proceedings of the 14th Python in Science Conference (SCIPY), Austin, TX, USA, 6–12 July 2015; pp. 18–25.
57. Bodlund, K. Alternative reference curves for evaluation of the impact sound insulation between dwellings. *J. Sound Vib.* **2017**, *116*, 173–181. [\[CrossRef\]](#)
58. Park, S.H.; Lee, P.J. Effects of floor impact noise on psychophysiological responses. *Build. Environ.* **2017**, *116*, 173–181. [\[CrossRef\]](#)
59. Deng, J.; Dong, W.; Socher, R.; Li, L.; Li, K.; Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Miami, FL, USA, 20–25 June 2009; pp. 248–255.
60. Pan, S.J.; Yang, Q. A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* **2009**, *22*, 1345–1359. [\[CrossRef\]](#)
61. Oquab, M.; Bottou, L.; Laptev, I.; Sivic, J. Learning and transferring mid-level image representations using convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Columbus, OH, USA, 23–28 June 2014; pp. 1717–1724.
62. Marmanis, D.; Datcu, M.; Esch, T.; Stilla, U. Deep learning earth observation classification using ImageNet pretrained networks. *IEEE Geosci. Remote Sens. Lett.* **2015**, *13*, 105–109. [\[CrossRef\]](#)

63. Zhang, S.; Qin, Y.; Sun, K.; Lin, Y. Few-Shot Audio Classification with Attentional Graph Neural Networks. In Proceedings of the INTERSPEECH, Graz, Austria, 15–19 September 2019; pp. 3649–3653.
64. He, K.; Zhang, X.; Ren, S.; Sun, J. Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 7–12 June 2015; pp. 1026–1034.
65. Aytar, Y.; Vondrick, C.; Torralba, A. SoundNet: Learning sound representations from unlabeled video. In Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS), Barcelona, Spain, 5–10 December 2016; pp. 892–900.
66. Bergstra, J.; Bengio, Y. Random search for hyper-parameter optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
67. Larochelle, H.; Erhan, D.; Courville, A.; Bergstra, J.; Bengio, Y. An empirical evaluation of deep architectures on problems with many factors of variation. In Proceedings of the 24th International Conference on Machine Learning (ICML), Corvallis, OR, USA, 20–24 June 2007; pp. 473–480.
68. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
69. Van der Maaten, L.; Hinton, G. Visualizing data using t-SNE. *J. Mach. Learn. Res.* **2008**, *9*, 2579–2605.